DOCUMENT RESUME

ED 264 260 TM 850 744

AUTHOR	Weiss, David J., Ed.
TITLE	Item Response Theory and Computerized Adaptive Testing Conference Proceedings (Wayzata, Minnesota, July 27-30, 1982).
INSTITUTION	Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY	Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
REPORT NO	ONR-85-1
PUB DATE	Apr 85
GRANT	N00014-82-G-0061
NOTE	384p.; For individual papers, see TM 850 745-746, TM 850 748-752, TM 850 754-763, ED 223 666, ED 243 901, and ED 249 281.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC16 Plus Postage.
DESCRIPTORS	Achievement Tests; *Adaptive Testing; Branching; *Computer Assisted Testing; *Computer Oriented Programs; Individual Testing; Item Analysis; *Latent Trait Theory; Mathematical Models; *Psychometrics; Sequential Approach; Statistical Analysis; Statistical Studies; Test Construction; Test Items; *Test Theory
IDENTIFIERS	Item Characteristic Function; Office of Naval Research; *Parameter Estimation of Sequential Testing; Robustness; University of Minnesota

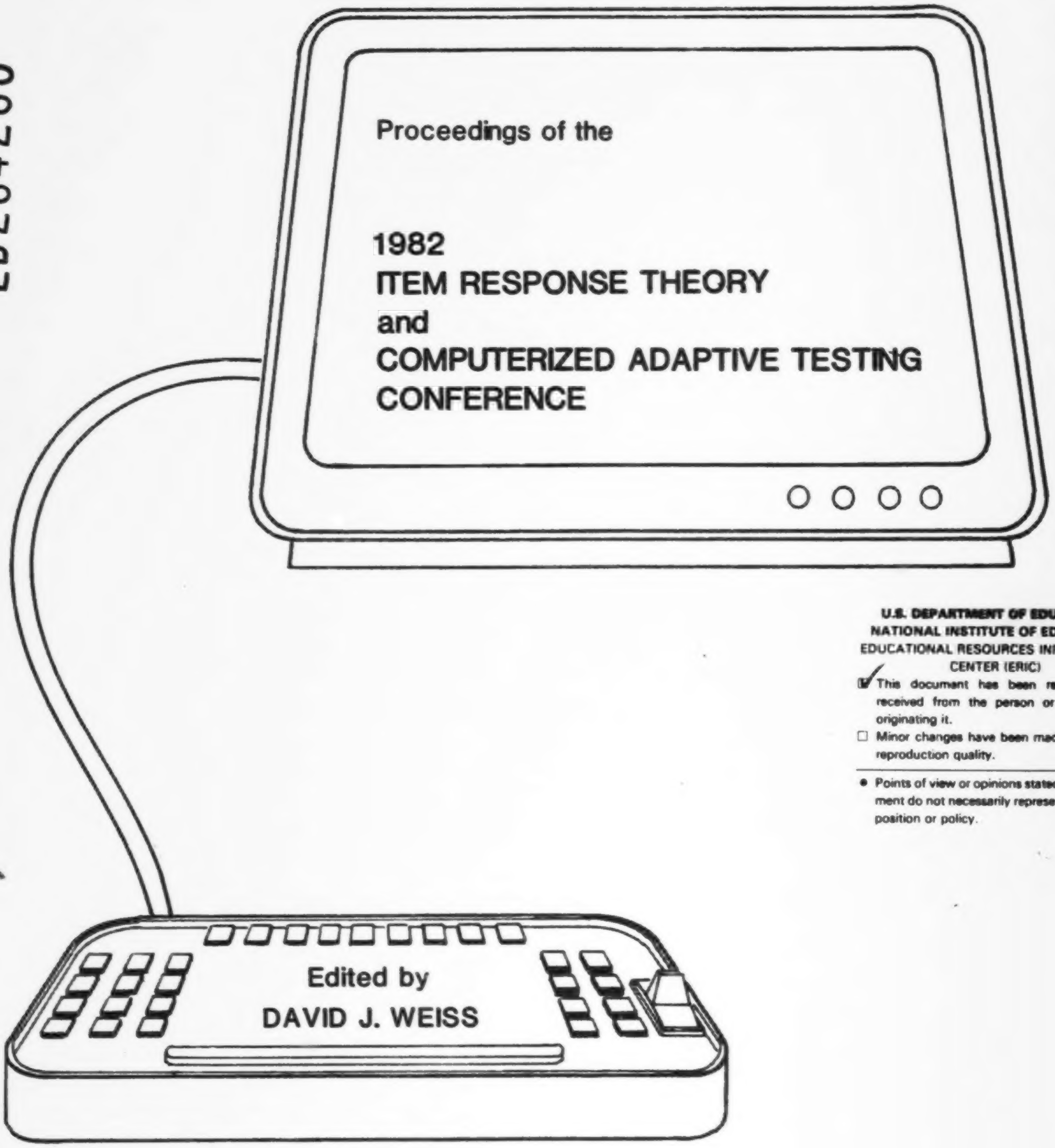
ABSTRACT

This report contains the Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. The papers and their discussions are organized into eight sessions: (1) "Developments in Latent Trait Theory," with papers by Fumiko Samejima and Michael V. Levine; (2) "Parameter Estimation," with papers by Frederic M. Lord and Marilyn S. Wingersky, David Thissen and Howard Wainer, and Charles Lewis; (3) "Multidimensional Item-Response Theory," with papers by Roderick P. McDonald, and Mark D. Reckase and Robert L. McKinley; (4) "Estimating Parameters with the E-M Algorithm," with papers by Robert K. Tsutakawa, and Robert J. Mislevy and R. Darrell Bock; (5) "Unidimensionality and Robustness," with papers by William Stout, Fritz Drasgow and Charles K. Parsons and Douglas H. Jones; (6) "Adaptive and Sequential Testing," with papers by David J. Weiss and Debra Suhadolnik and R. A. Weitzman; (7) "Latent Trait Models for Special Applications," with papers by Susan Embretson (Whitely) and Kikumi K. Tatsuoka; (8) "Applications of Computerized Adaptive Testing," with papers by James R. McBride and J. B. Sympson, J. B. Sympson and Loralee Hartmann, Wolfgang Wildgrube, and C. David Vale. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

ED264260



Proceedings of the

1982
ITEM RESPONSE THEORY
and
COMPUTERIZED ADAPTIVE TESTING
CONFERENCE

Edited by
DAVID J. WEISS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 744

PROCEEDINGS OF THE
1982
ITEM RESPONSE THEORY
AND
COMPUTERIZED ADAPTIVE TESTING
CONFERENCE

held at the
Spring Hill Conference Center
Wayzata, Minnesota
July 27-30, 1982

EDITED BY
DAVID J. WEISS

COMPUTERIZED ADAPTIVE TESTING LABORATORY
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS MN 55455
APRIL 1985

Prepared under Department of the Navy Grant No. N00014-82-G-0061
issued by the Personnel and Training Research Programs,
Psychological Sciences Division, Office of Naval Research

Approved for public release, distribution unlimited.
The United States Government has a royalty-free license
throughout the world in all copyrightable material contained herein.

CONTENTS

Developments in Latent Trait Theory

Development and Application of Methods for Estimating Operating Characteristics of Discrete Test Item Responses without Assuming Any Mathematical Form
Fumiko Samejima 1

Discussion
Roderick P. McDonald 39

The Trait in Latent Trait Theory
Michael V. Levine 41

Discussion
Robert J. Mislevy 66

Parameter Estimation

Sampling Variances and Covariances of Parameter Estimates in Item Response Theory
Frederic M. Lord and Marilyn S. Wingersky 69

Confidence Envelopes for Item Response Functions
David Thissen and Howard Wainer 89

Discussion
Michael V. Levine 101

Developments in Nonparametric Ability Estimation
Charles Lewis 105

Discussion
Robert K. Tsutakawa 123

Multidimensional Item Response Theory

Unidimensional and Multidimensional Models for Item Response Theory
Roderick P. McDonald 127

Discussion
Fumiko Samejima 149

Some Latent Trait Theory in a Multidimensional Latent Space
Mark D. Reckase and Robert L. McKinley 151

Discussion
Fritz Drasgow 178

Estimating Parameters with the E-M Algorithm

Estimation of Item Parameters and the GEM Algorithm
Robert K. Tsutakawa 180

Implementation of the EM Algorithm in the Estimation of Item Parameters: The BILOG Computer Program
Robert J. Mislevy and R. Darrell Bock 189

Discussion
Charles Lewis 203

CONTENTS, continued

Unidimensionality and Robustness

A Statistical Procedure for Assessing Test Dimensionality
William Stout 210

Application of Unidimensional Item Response Theory Models to Multidimensional Data
Fritz Drasgow and Charles K. Parsons 218

Tools of Robustness for Item Response Theory
Douglas H. Jones 233

Discussion
David Thissen 244

Adaptive and Sequential Testing

Robustness of Adaptive Testing to Multidimensionality
David J. Weiss and Debra Suhadolnik 248

Use of Sequential Testing to Prescreen Prospective Entrants into Military Service
R. A. Weitzman 281

Discussion
Mark D. Reckase 290

Latent Trait Models for Special Applications

Component Latent Trait Models for Test Design
Susan Embretson (Whitely) 295

Discussion: Multidimensional Latent Traits and Plausible Assumptions of Cognitive Psychology
Douglas H. Jones 317

A Latent Trait Model for Interpreting Misconceptions in Procedural Domains
Kikumi K. Tatsuoka 322

Discussion
Susan Embretson (Whitely) 340

Applications of Computerized Adaptive Testing

The Computerized Adaptive Testing System Development Project
James R. McBride and J. B. Sympson 342

Item Calibrations for Computerized Adaptive Testing (CAT) Experimental Item Pools
J. B. Sympson and Loralee Hartmann 350

Computerized Testing in the German Federal Armed Forces (FAF): Empirical Approaches
Wolfgang Wildgrube 353

Design of a Microcomputer-Based Adaptive Testing System
C. David Vale 360

DEVELOPMENT AND APPLICATION OF METHODS FOR ESTIMATING OPERATING CHARACTERISTICS OF DISCRETE TEST ITEM RESPONSES WITHOUT ASSUMING ANY MATHEMATICAL FORM

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE

In latent trait theory the latent space, or space of the hypothetical construct, is usually represented by some unidimensional or multidimensional continuum of real numbers. Many researchers have concentrated their effort on unidimensional spaces, although there have been some models developed in the multidimensional latent space (Samejima, 1974a, 1974b), and this will be an important future orientation of research. Like the latent space, the item response can either be treated as a discrete variable or as a continuous variable. Although most researchers have solely treated discrete responses, there are certain papers which have dealt with continuous responses (Samejima, 1973a, 1974a).

Latent trait theory relates the item response to the latent trait in terms of the operating density characteristic when the item response is continuous, and by means of the operating characteristic when the item response is discrete. By the operating density characteristic of the continuous item response is meant the conditional density function of a particular item response, given the latent trait. This operating density characteristic is replaced by the operating characteristic of the item response on the discrete response level, which is the conditional probability function of the item response, given the latent trait. The discrete response level includes the nominal response level, the graded response level, and the dichotomous response level (Samejima, 1972).

On the nominal response level, response categories are not explicitly ordered. Using information given by the incorrect as well as the correct answer of the multiple-choice test item provides a good example of the nominal response level. Bock (1972) proposed a multinomial response model, which is applicable to the multiple-choice test item when the effect of random guessing is negligibly small. Samejima (1979b) proposed a family of models for the multiple-choice test item, which accounts both for the information provided by the incorrect answers and for the effect of random guessing.

The graded response level is further categorized into the homogeneous and heterogeneous cases. Samejima (1969, 1972) has proposed the normal ogive model and the logistic model developed for the homogeneous case of the graded response level. In a special case of the graded response level where the item response is binary, e.g., correct or incorrect in ability or achievement measurement,

Table 1

Notation

θ : unidimensional latent trait

g : item of Old Test (=1, 2, ..., n)

k_g : discrete item response to item g

$P_{k_g}(\theta)$: operating characteristic of k_g

$A_{k_g}(\theta)$: basic function of k_g

V : response pattern or vector of n discrete item responses

$L_V(\theta)$: likelihood function of θ given response pattern

$P_V(\theta)$: operating characteristic response pattern V

$I_{k_g}(\theta)$: item response information function of discrete response k_g

$I_g(\theta)$: item information function of item g

$I_V(\theta)$: response pattern information function of V

$I(\theta)$: test information function of Old Test

$\hat{\theta}_V$, $\hat{\theta}$: maximum likelihood estimate of θ

ε : error of maximum likelihood estimate $\hat{\theta}$

x_g : graded item response to item g (= 0, 1, ..., m)
(special case of k_g)

$P_{x_g}(\theta)$: operating characteristic of x_g

a_g : discrimination parameter of item g

b_{x_g} : difficulty parameter of x_g

$P_g(\theta)$: item characteristic function of item g

b_g : difficulty parameter of item g when g is binary

$\underline{\theta}$, $\overline{\theta}$: lower and upper endpoints of arbitrarily chosen subinterval of θ

$\underline{\underline{\theta}}$, $\overline{\overline{\theta}}$: lower and upper endpoints of the interval of θ for which $P_g(\theta)$ of
Type A is strictly increasing in θ

$\sigma(\theta)$: standard error of maximum likelihood estimate $\hat{\theta}_V$ given θ

$f(\theta)$: density function of latent trait θ

τ : transformed latent trait, which is strictly increasing in θ

$P^*_{k_g}(\tau)$: operating characteristic of k_g defined on τ

$P^*_V(\tau)$: operating characteristic of V defined on τ

- continued on next page -

Table 1 (continued)

Notation

$I^*_{k_g}(\tau)$: item response information function defined on τ

$I^*_g(\tau)$: item information function defined on τ

$I^*_V(\tau)$: response pattern information function defined on τ

$I^*(\tau)$: test information function defined on τ

$\hat{\tau}_V$, $\hat{\tau}$: maximum likelihood estimate of τ based on V

$\underline{\tau}$, $\overline{\tau}$: lower and upper endpoints of the interval corresponding to $(\underline{\theta},\overline{\theta})$

τ_0 : value of τ corresponding to the origin of θ

$f^*(\tau)$: density function of transformed latent trait τ

C : constant square root of test information defined on τ

α_k : $(k = 0, 1, \ldots, m)$ coefficients of polynomial approximating the square root of the test information function of θ

α^*_k : $(k = 0, 1, \ldots, m+1)$ coefficients of polynomial transforming θ to τ

ε^* : error of maximum likelihood estimate $\hat{\tau}$

$\dot{\varepsilon}(\hat{\tau})$: sample linear regression of ε^* on $\hat{\tau}$

$g(\hat{\tau})$: density function of $\hat{\tau}$

s : examinee $(= 1, 2, \ldots, N)$

$\hat{\tau}_s$: maximum likelihood estimate of $\hat{\tau}$ assigned to examinee $\underline{s}$

h : test item whose operating characteristics are to be discovered, or "unknown test item"

k_h : discrete item response to unknown test item $\underline{h}$

$P_{k_h}(\theta)$: operating characteristic of k_h

$f_{k_h}(\theta)$: density function of θ for the subgroup of examinees who share the discrete item responses k_h to item $\underline{h}$

N_{k_h} : number of examinees who share the discrete item response k_h to item $\underline{h}$ in the sample

$g_{k_h}(\hat{\tau})$: density function of $\hat{\tau}$ for the subgroup of examinees who share the discrete item response k_h to item $\underline{h}$

$\phi_{k_h}(\tau|\hat{\tau})$: conditional density function of τ, given $\hat{\tau}$

$\xi_{k_h}(\tau,\hat{\tau})$: joint density function of τ and $\hat{\tau}$

$w(\hat{\tau}_s)$: weight assigned to $\hat{\tau}_s$

$p(s\varepsilon k_h)$: probability with which the examinee $\underline{s}$ belongs to the item response subgroup k_h

Table 2
Fundamental Formulae

$$V = (k_1, k_2, \ldots, k_g, \ldots, k_n)' \qquad [1]$$

$$P_V(\theta) = \prod_{k_g \varepsilon V} P_{k_g}(\theta) \quad \text{(by local independence)} \qquad [2]$$

$$\frac{\partial}{\partial\theta} \log L_V(\theta) = \frac{\partial}{\partial\theta} \log P_V(\theta) = \sum_{k_g \varepsilon V} A_{k_g}(\theta) = 0$$

$$\text{(likelihood equation)} \qquad [3]$$

$$A_{k_g}(\theta) = \frac{\partial}{\partial\theta} \log P_{k_g}(\theta) \qquad [4]$$

$$I_{k_g}(\theta) = -\frac{\partial}{\partial\theta} A_{k_g}(\theta) \qquad [5]$$

$$I_g(\theta) = E[I_{k_g}(\theta) \mid \theta] = \sum_{k_g} I_{k_g}(\theta)\, P_{k_g}(\theta)$$

$$= \sum_{k_g} [\frac{\partial^2}{\partial\theta^2} P_{k_g}(\theta)]^2 [P_{k_g}(\theta)]^{-1} \qquad [6]$$

$$I_V(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_V(\theta) = \sum_{k_g \varepsilon V} I_{k_g}(\theta) \qquad [7]$$

$$I(\theta) = E[I_V(\theta) \mid \theta] = \sum_V I_V(\theta)\, P_V(\theta) \qquad [8]$$

$$I(\theta) = \sum_{g=1}^{n} I_g(\theta) \quad \text{(derived from Equations 6, 7, and 8)} \qquad [9]$$

$$\tau = \tau(\theta) \qquad [10]$$

$$P^*_{k_g}[\tau(\theta)] = P_{k_g}(\theta) \qquad [11]$$

$$P^*_V[\tau(\theta)] = P_V(\theta) \qquad [12]$$

- continued on the next page -

Table 2 (continued)
Fundamental Formulae

$$I^*_{k_g}(\tau) = I_{k_g}(\theta)\,[\frac{d\theta}{d\tau}]^2 - \frac{\partial}{\partial\theta}\log P_{k_g}(\theta)\cdot\frac{d^2\theta}{d\tau^2}$$ [13]

$$I^*_g(\tau) = I_g(\theta)\,[\frac{d\theta}{d\tau}]^2$$ [14]

$$I^*_V(\tau) = I_V(\theta)\,[\frac{d\theta}{d\tau}]^2 - \frac{\partial}{\partial\theta}\log P_V(\theta)\cdot\frac{d^2\theta}{d\tau^2}$$ [15]

$$I^*(\tau) = I(\theta)\,[\frac{d\theta}{d\tau}]^2$$ [16]

$$\begin{cases} \underline{\tau} = \tau(\underline{\theta}) \\ \bar{\tau} = \tau(\bar{\theta}) \end{cases}$$ [17]

$$\int_{\underline{\tau}}^{\bar{\tau}} [I^*_g(\tau)]^{1/2}\,d\tau = \int_{\underline{\theta}}^{\bar{\theta}} [I_g(\theta)]^{1/2}\,d\theta$$

(constancy of square root of item information) [18]

$$\int_{\underline{\tau}}^{\bar{\tau}} [I^*(\tau)]^{1/2}\,d\tau = \int_{\underline{\theta}}^{\bar{\theta}} [I(\theta)]^{1/2}\,d\theta$$

(constancy of square root of test information) [19]

---

positive or negative in attitude measurement, and so forth, the operating characteristic of the positive response is called the item characteristic function (Lord & Novick, 1968, chap. 16). On this dichotomous response level, mathematical models such as the Rasch (1960) model, the normal ogive model (Lord, 1952; Tucker, 1951), the logistic model and its 3-parameter version (Birnbaum, 1968) have been proposed and used.

The present paper is concerned solely with the unidimensional latent space and the discrete response level. Samejima (1977c, 1977e, 1978a, 1978b, 1978c, 1978d, 1978e, 1978f, 1980a, 1980b) has developed a series of methods and approaches for estimating the operating characteristics of discrete item responses without assuming any mathematical form. The outline of those methods and approaches will be introduced in the present paper. The direct estimation of the operating characteristics of discrete item responses has also been attempted by Lord (1969, 1970), in which the estimation of the true test score distribution of each subgroup of examinees plays an essential role. Levine (1980) has also developed a method for a similar purpose, in which he has utilized eigenfunctions effectively. Some observations concerning model validation, scale construction, and information loss will also be made.

Notation and Fundamental Formulae

Table 1 presents the set of basic symbols used in the present study. Assume that there is a set of n items, or Old Test, whose operating characteristics are known, and let g denote an arbitary item of the Old Test. An "unknown" test item i.e., any test item whose operating characteristics are to be discovered, is denoted by h, to distinguish it from item g. Also assume that the Old Test has a substantially large amount of test information at any point of the interval of θ of interest.

Table 2 presents the set of fundamental formulae which relate these mathematical symbols with one another. They are the basis of the present study.

Asymptotic Properties of the Maximum Likelihood Estimate and Transformation of Latent Trait

The well-known property of the maximum likelihood estimate--under a general condition its distribution converges to normality--is fully utilized in the present study. The proof of this theorem is given in many statistics books when the observations are taken from identical distributions (e.g., Kendall & Stuart, 1961). It can easily be expanded to the general case where the test items are not necessarily equivalent (Samejima, 1976).

Let ε be the error in the maximum likelihood estimate, $\hat{\theta}_V$, of the latent trait θ, so that

$$\hat{\theta}_V = \theta + \varepsilon \ . \qquad [20]$$

Thus, $\hat{\theta}_V$ is asymptotically and conditionally unbiased and normally distributed with $\sigma(\theta)$ as the second parameter, such that

$$E(\hat{\theta}_V | \theta) \simeq \theta \qquad [21]$$

and

$$\sigma(\theta) \simeq [I(\theta)]^{-\frac{1}{2}} \ . \qquad [22]$$

The speed of convergence of this conditional distribution to normality has been observed by monte carlo studies (Samejima, 1976, 1977a, 1977b, 1977c, 1979a). In three studies, except for the last one in which the Constant Information Model was adopted, either the normal ogive model on the dichotomous response level or the one on the graded response level was used. The item characteristic function, $P_g(\theta)$, of the first model is given by

$$P_g(\theta) = [2\pi]^{-\frac{1}{2}} \int_{-\infty}^{a_g(\theta-b_g)} e^{-u^2/2} \, du \ , \qquad [23]$$

where a_g (>0) and b_g are the item discrimination and difficulty parameters, respectively. For the operating characteristic of the graded response x_g (= 0, 1, ..., m_g) in the second model and

$$P_{x_g}(\theta) = [2\pi]^{-1/2} \int_{a_g(\theta-b_{x_g+1})}^{a_g(\theta-b_{x_g})} \qquad [24]$$

where

$$-\infty = b_0 < b_1 < \ldots < b_{m_g} < b_{m_g+1} = \infty \ . \qquad [25]$$

These different monte carlo studies have shown that the speed of convergence is high, i.e., the conditional distribution of $\hat{\theta}_V$, given θ, can be approximated by $N(\theta, I(\theta)^{-1/2})$ even when the number of test items is relatively small. The benefit of this finding will be enhanced if the test information function, $I(\theta)$, assumes a constant value for the interval of interest, for it will simplify the subsequent mathematics. In computerized adaptive testing the constant standard error of estimation can be obtained easily for the interval of θ of interest (Samejima, 1977b). This can be accomplished by using a set amount of test information as the criterion for terminating the presentation of new test items for each individual examinee. In so doing, weakly parallel tests (Samejima, 1977d) are produced in each session, with the same accuracy of estimation everywhere within the interval of θ that is of interest.

When this is not the case, the latent trait θ can be transformed to τ, which provides constant test information (cf. Samejima, 1980a). Let C be the desired constant value of the square root of the test information function for the interval $[\underline{\tau},\overline{\tau}]$. Then, the transformation of θ to τ is given by

$$\tau(\theta) = C^{-1} \int_{-\infty}^{\theta} [I(u)]^{1/2}\, du + \tau_0 \ , \qquad [26]$$

where τ_0 is an arbitrary constant, or the value of τ which corresponds to the origin of θ. Note, however, that in order to approximate the conditional distribution of the maximum likelihood estimate $\hat{\tau}_V$, given the latent trait τ, it is necessary that in spite of its irregularity, the amount of test information defined for the original latent trait θ be substantially large throughout the interval of θ of interest. By virtue of the transformation-free character of the maximum likelihood estimate, the new maximum likelihood estimate, $\hat{\tau}_V$, can be obtained through the same transformation, such that

$$\hat{\tau}_V = \tau(\hat{\theta}_V) \ . \qquad [27]$$

The transformation of θ to τ, which is given by Equation 26, may be rather complicated, for it includes the integration of the square root of the test information function I(θ), given by Equation 9 (see Table 2). The process will be largely simplified if the square root of the test information function is approximated by a polynomial obtained by the method of moments (Elderton & Johnson, 1969). It has been observed (Samejima & Livingston, 1979) that the polynomial of a specified degree, m, which is obtained by the method of moments, provides the least squares solution. By the use of the method of moments, rounding error is also reduced, for it does not include the inversion of an ill-conditioned matrix. With such a polynomial, for the square root of the test information function

$$[I(\theta)]^{1/2} \doteq \sum_{k=0}^{m} \alpha_k \theta^k . \qquad [28]$$

can be written. Substituting Equation 28 into Equation 27 gives

$$\tau(\theta) \doteq C^{-1} \sum_{k=0}^{m} \alpha_k (k+1)^{-1} \theta^{k+1} + \tau_0$$

$$= \sum_{k=0}^{m+1} \alpha_k^* \theta^k , \qquad [29]$$

where

$$\alpha_k^* \begin{cases} = \tau_0 & k = 0 \\ = (Ck)^{-1} \alpha_{k-1} & k = 1,2, \ldots, m+1 . \end{cases} \qquad [30]$$

Approach to the Joint Density Function of τ and $\hat{\tau}_V$

When the operating characteristics of the test items of the Old Test satisfy the unique maximum condition (Samejima, 1969, 1972), the maximum likelihood estimate of the latent trait θ can be obtained for each of those examinees to whom the Old Test was administered. Perhaps the most naive way of obtaining the estimated operating characteristics of a discrete item response is to take the simple ratio of the frequency of the examinees who share the response to the total frequency, by setting small subintervals of $\hat{\theta}_V$. This method has certain serious theoretical problems, however, in addition to the fact that it requires a large sample size and, in general, a substantial amount of computer time in obtaining $\hat{\theta}_V$. Generally speaking, the regression of θ on $\hat{\theta}_V$ is not even linear. Thus, even if the maximum likelihood estimate $\hat{\theta}_V$ is approximately conditionally unbiased, given the latent trait θ and constant test information for the interval of θ of interest, the resultant estimated operating characteristic will be stretched unevenly. For the expectation of $\hat{\theta}_V$

$$E(\hat{\theta}_V) = E(\theta) \ , \qquad [31]$$

when the maximum likelihood estimate $\hat{\theta}_V$ is approximately conditionally unbiased, given θ. The general formula for the mth moment of $\hat{\theta}_V$ is given by

$$E[\hat{\theta}_V - E(\hat{\theta}_V)]^m = \sum_{r=0}^{m} \binom{m}{r} E[\{\theta - E(\theta)\}^{m-r} \, E\{(\hat{\theta}_V - \theta)^r | \theta\}] \qquad [32]$$

(cf. Samejima, 1977c). From Equation 32

$$\mathrm{Var}(\hat{\theta}_V) = \mathrm{Var}(\theta) + E[\mathrm{Var}(\hat{\theta}_V|\theta)] \geq \mathrm{Var}(\theta) \ , \qquad [33]$$

is obtained for the variance of $\hat{\theta}_V$, ignoring the covariance term for the case where the conditional unbiasedness of $\hat{\theta}_V$ approximately holds. For these reasons, it is desirable to discover methods which ameliorate these deficiencies.

The group of examinees of the sample can be categorized into subgroups depending upon their item scores, k_h, to the "unknown" item h. If the density function, $\tilde{f}_{k_h}(\theta)$, of each discrete item score k_h is estimated, then the estimated operating characteristics, $\tilde{P}_{k_h}(\theta)$, can be represented by

$$\tilde{P}_{k_h}(\theta) = N_{k_h} \cdot \tilde{f}_{k_h}(\theta) \ [\sum_j N_j \cdot \tilde{f}_j(\theta)]^{-1} \ , \qquad [34]$$

where N_{k_h} is the number of examinees who belong to the item response category k_h of item h and $\tilde{f}_{k_h}(\theta)$ is the estimated density. Asssume that the Old Test has a sufficiently large amount of test information at any point in the interval of θ of interest, and θ is transformed to τ, which has a constant test information throughout the corresponding interval of τ. Then

$$\tau_V = \tau + \varepsilon^* \ , \qquad [35]$$

where ε^* denotes the error in estimating $\hat{\tau}_V$.

In the Normal Approximation Method, a bivariate normal distribution is assumed for the joint distribution of $\hat{\tau}_V$ and τ for each subpopulation of examinees who share the same discrete item response k_h to the unknown item h. For simplicity, hereafter, the subscript V will be dropped from $\hat{\tau}_V$. The sample linear regression, $\dot{\varepsilon}^*(\hat{\tau})$ of the error ε^* on the maximum likelihood estimate $\hat{\tau}$ is given by

$$\dot{\varepsilon}^*(\hat{\tau}) = \alpha + \beta\hat{\tau} \ , \qquad [36]$$

where

$$\alpha = -C^{-2} \, E(\hat{\tau}) \ [\mathrm{Var}(\hat{\tau})]^{-1} \qquad [37]$$

and

$$\beta = C^{-2}\ [\mathrm{Var}(\hat{\tau})\}^{-1}]\ . \qquad [38]$$

For the conditional variance of ε^*, given $\hat{\tau}$,

$$\mathrm{Var}(\overset{*}{\varepsilon}|\hat{\tau}) \doteq C^{-2}[1 - C^{-2}\{\mathrm{Var}(\hat{\tau})\}^{-1}]\ . \qquad [39]$$

can also be written. Thus, the approximate joint distribution of ε^* and $\hat{\tau}$ are completely specified through Equations 36 through 39 for each subpopulation of examinees. Using the monte carlo method, as many error scores as desired can be produced for each value of $\hat{\tau}$; subtracting each value of ε^* thus produced from $\hat{\tau}$, the value of τ is obtained, which will be denoted by $\tilde{\tau}$. The resulting set of τ^* for each item response subgroup k_h will be appropriately categorized into small subintervals to provide the approximate product of the interval width of θ and the frequency $N_{k_h}\tilde{f}_{k_h}(\theta)$ in Equation 34, by virtue of the one-to-one correspondence between θ and τ; therefore, it can be used in both the numerator and denominator of Equation 34.

A hypothetical test, called Test A, which consists of 35 graded items, was used as the Old Test. Each item of Test A follows the normal ogive model on the graded response level, whose operating characteristics, $P_{x_g}(\theta)$, is given by Equation 35. The item discrimination parameter, $a_g (> 0)$, and the set of m_g (= 2) item difficulty parameters, b_{x_g}, of each of the 35 test items of Test A are presented in Table 3. The amount of test information of Test A is approximately 21.63 throughout the interval of θ (-3.0, 3.0); therefore, the standard error of estimate is approximately .215 for this interval of θ (cf. Samejima, 1977c). Five hundred hypothetical examinees whose positions on the latent trait θ are at 100 equally spaced points of θ starting with -2.475 and ending with 2.475, with five examinees placed at each position, were used. For Test A, τ can be set equal to θ, and the transformation of the latent trait is not necessary. These are 10 "unknown" binary items, each of which follows the normal ogive model on the dichotomous response level (whose item characteristic function is given by Equation 23, and whose two item parameters, a_g and b_g, are presented in Table 4).

Figure 1 presents the results obtained by the Normal Approximation Method (Samejima, 1977c), choosing Item 6 as an example. The result obtained by producing only one value of $\tilde{\theta}$ for each $\tilde{\theta}_V$ is plotted by hollow circles (N = 500); the two sets of results obtained by producing five values of $\tilde{\theta}$ for each are plotted by solid triangles (N = 2,500) and X's (N = 2,500), respectively; the result obtained by combining these two sets of five values of $\tilde{\theta}$ for each $\tilde{\theta}$ is plotted by hollow squares (N = 5,000). It can be seen that the accuracy of the estimation of the operating characteristic becomes higher as a larger number of $\tilde{\theta}$ are produced. Similar results were obtained for the other nine unknown test items, i.e., Items 1 through 5 and 7 through 10.

The Normal Approximation Method approximates the joint distribution of τ and $\hat{\tau}$ for each subpopulation by a bivariate normal distribution. A similar approach is possible with somewhat different rationale using the marginal density

Table 3
Item Discrimination Parameter
a_g and Two Item Difficulty
Parameters b_{xg} for $x_g=1$ and
$x_g=2$ of Each of the 35
Graded Items of Test A

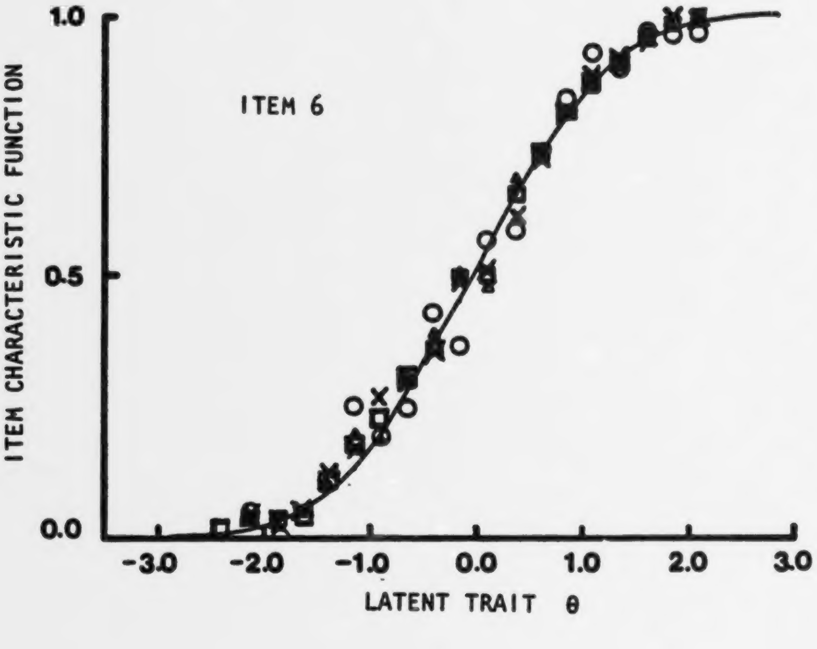
Item g	a_g	b_1	b_2
1	1.8	-4.75	-3.75
2	1.9	-4.50	-3.50
3	2.0	-4.25	-3.25
4	1.5	-4.00	-3.00
5	1.6	-3.75	-2.75
6	1.4	-3.50	-2.50
7	1.9	-3.00	-2.00
8	1.8	-3.00	-2.00
9	1.6	-2.75	-1.75
10	2.0	-2.50	-1.50
11	1.5	-2.25	-1.25
12	1.7	-2.00	-1.00
13	1.5	-1.75	-0.75
14	1.4	-1.50	-0.50
15	2.0	-1.25	-0.25
16	1.6	-1.00	0.00
17	1.8	-0.75	0.25
18	1.7	-0.50	0.50
19	1.9	-0.25	0.75
20	1.7	0.00	1.00
21	1.5	0.25	1.25
22	1.8	0.50	1.50
23	1.4	0.75	1.75
24	1.9	1.00	2.00
25	2.0	1.25	2.25
26	1.6	1.50	2.50
27	1.7	1.75	2.75
28	1.4	2.00	3.00
29	1.9	2.25	3.25
30	1.6	2.50	3.50
31	1.5	2.75	3.75
32	1.7	3.00	4.00
33	1.8	3.25	4.25
34	2.0	3.50	4.50
35	1.4	3.75	4.75

function of $\hat{\tau}$ (Samejima, 1978f). Let $g_{k_h}(\hat{\tau})$ denote the density function of the maximum likelihood estimate $\hat{\tau}$ and $\phi_{k_h}(\tau|\hat{\tau})$ be the conditional density function of τ, given $\hat{\tau}$, for each subpopulation of examinees who share the same discrete item response k_h. For the joint density function, $\xi_{k_h}(\tau,\hat{\tau})$, of τ and $\hat{\tau}$

Table 4
Item Discrimination
Parameter a_h and
Item Difficulty
Parameter b_h of
Each "Unknown"
Test Item h

Item h	a	b
1	1.5	-2.5
2	1.0	-2.0
3	2.5	-1.5
4	1.0	-1.0
5	1.5	-0.5
6	1.0	0.0
7	2.0	0.5
8	1.0	1.0
9	2.0	1.5
10	1.0	2.0

Figure 1
Estimated Item Characteristic Functions of Item 6 Based upon 500 $\tilde{\theta}$'s (Hollow Circles), 2,500 $\tilde{\theta}$'s (Solid Triangles), 2,500 $\tilde{\theta}$'s (Crosses), and 5,000's $\tilde{\theta}$'s (Hollow Squares), Obtained by the Normal Approximation Method, Using Test A as the Old Test



$$\xi_{k_h}(\tau,\hat{\tau}) = \phi_{k_h}(\tau|\hat{\tau})\; g_{k_h}(\hat{\tau}) \;. \qquad [40]$$

can be written. From these joint density functions for the separate subpopulations and the frequencies of examinees, N_{k_h}, are obtained for the estimated operating characteristic

$$\tilde{P}_{k_h}[\theta(\tau)] = N_{k_h}\int_{-\infty}^{\infty} \xi_{k_h}(\tau,\hat{\tau})d\hat{\tau}\;[\sum_{j} N_j \int_{-\infty}^{\infty} \xi_j(\tau,\hat{\tau})d\hat{\tau}]^{-1} \;. \qquad [41]$$

This approach is called the Bivariate P.D.F. Approach.

The density function, $g_{k_h}(\hat{\tau})$, in Equation 40 must be estimated from the observable set of maximum likelihood estimates. This can be done by fitting a polynomial of an appropriate degree to the set of maximum likelihood estimates by the method of moments (cf. Samejima & Livingston, 1979).

The conditional density, $\phi_{k_h}(\tau|\hat{\tau})$, in Equation 40 can be estimated by using the conditional moments of τ, given $\hat{\tau}$. It has been shown (Samejima, 1977e) that if an estimate of the latent trait is conditionally unbiased and its conditional distribution is normal with a constant second parameter, the conditional moments of the latent trait, given its maximum likelihood estimate, will be obtained solely from the density function of the estimate and the constant second parameter of the normal distribution. The conditional expectation of τ, given $\hat{\tau}$, and the second through fourth conditional moments of τ, given $\hat{\tau}$, are given by the following formulae:

$$E(\tau|\hat{\tau}) = \hat{\tau} + C^{-2}\frac{d}{d\hat{\tau}} \log g(\tau) \qquad [42]$$

$$Var(\tau|\hat{\tau}) = C^{-2}[1 + C^{-2}\frac{d^2}{d\tau^2} \log g(\hat{\tau})] \qquad [43]$$

$$E[\{\tau - E(\tau|\hat{\tau})\}^3|\hat{\tau}] = C^{-6}\left[\frac{d^3}{d\tau^3} \log g(\hat{\tau})\right] \qquad [44]$$

$$E[\{\tau - E(\tau|\hat{\tau})\}^4|\hat{\tau}] = C^{-4}\Big[3 + 6C^{-2}\Big\{\frac{d^2}{d\tau^2} \log g(\hat{\tau})\Big\} + 3C^{-4}\Big\{\frac{d^2}{d\tau^2} \log g(\hat{\tau})\Big\}^2 + C^{-4}\Big\{\frac{d^4}{d\tau^4} \log g(\hat{\tau})\Big\}\Big] \qquad [45]$$

Let κ denote Pearson's criterion. Then,

$$\kappa = \beta_1(\beta_2 + 3)^2[4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)]^{-1} \qquad [46]$$

where

$$\beta_1 = (E[\{\tau - E(\tau|\hat{\tau})\}^3|\hat{\tau}])^2(\mathrm{Var}[\tau|\hat{\tau}])^{-3} \qquad [47]$$

and

$$\beta_2 = (E[\{\tau - E(\tau|\hat{\tau})\}^4|\hat{\tau}])(\mathrm{Var}[\tau|\hat{\tau}])^{-2} . \qquad [48]$$

The value of this criterion points to a unique Pearson Type distribution (Elderton & Johnson, 1969) which approximates the conditional density, $\phi_{k_h}(\tau|\hat{\tau})$, in Equation 40. This method will be called the Pearson System Method (Samejima, 1978b).

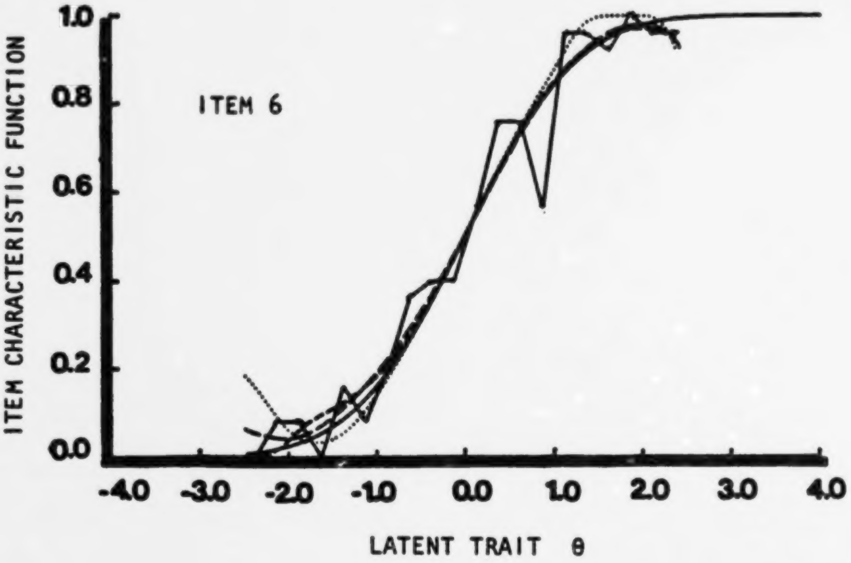
To avoid the increased inaccuracies of estimation of conditional moments of higher orders, it is sometimes suitable to use only the first and second conditional moments in approximating the conditional density, $\phi_{k_h}(\tau|\hat{\tau})$. The Normal Approach Method (Samejima, 1978b), which adopts a normal density function for the estimate of $\phi_{k_h}(\tau|\hat{\tau})$, is such a method, although each estimated conditional density function is forced to be symmetric. For added flexibility, a beta density function can also be used for the estimate of $\phi_{k_h}(\tau|\hat{\tau})$, whose two parameters, the lower and the upper endpoints of the interval for which the density is positive, are given a priori. Thus, as in the Normal Approach Method, only the first two conditional moments are needed to estimate the remaining two parameters of the beta distribution. This method will be called the Two-Parameter Beta Method (Samejima, 1977e, 1978a).

Figure 2 illustrates the results obtained by using the Bivariate P.D.F. Approach combined with the Normal Approach Method, with the same set of simulated data that was introduced earlier. The unknown test item illustrated here is, again, Item 6. Three different degrees, i.e., Degrees 3, 4 and 5, were used for the polynomials that were obtained by the method of moments to provide the estimated density function, $g_{k_h}(\hat{\theta})$, of each discrete item response k_h.

It can be seen in this example that the result of the Degree 3 case does not fit the theoretical item characteristic function as well as those of the Degree 4 and 5 cases, although it is still fairly close to the theoretical curve. This is not a general tendency, however, and those curves of the Degree 3 case fit just as closely to the corresponding theoretical item characteristic functions as those of the Degree 4 and 5 cases for most of the other nine unknown test items.

Comparison of these three resultant estimated item characteristic functions to the stepwise frequency ratios of the subgroup of examinees who answered Item 6 correctly to the total subgroup, which are based upon the true positions of the five hundred hypothetical examinees on the latent trait θ, indicates that all the three results are much closer to the theoretical item characteristic function than the frequency ratios.

Figure 2
Estimated Item Characteristic Functions of Item 6 for the Degree 3 (Dotted Curve), Degree 4 (Short Dashed Curve), and Degree 5 (Long Dashed Curve) Cases of the Bivariate P.D.F. Approach with the Normal Approach Method Using Test A as the Old Test, and the Theoretical Item Characteristic Function (Smooth Solid Curve) and the Frequency Ratios of θ (Jagged Solid Curve)



Conditional P.D.F. Approach

When there is a large number of unknown test items, the computing time required in the Normal Approximation Method or in the Bivariate P.D.F. Approach will be substantially large, since the joint density function must be approximated for each discrete item response to each unknown test item. In contrast, when it is desirable to deal with more than one unknown item together, the Conditional P.D.F. Approach is much simpler. This approach is further categorized into three procedures, i.e., the Simple Sum Procedure (Samejima, 1978a), the Weighted Sum Procedure (Samejima, 1978d), and the Proportioned Sum Procedure (Samejima, 1978e). Let $\hat{\tau}_s$ be the maximum likelihood estimate of the latent trait assigned to the examinee denoted s, and $w(\hat{\tau}_s)$ denote an appropriately chosen weight for $\hat{\tau}_s$. In the Weighted Sum Procedure the estimated operating characteristic of the discrete item response k_h to the unknown test item h is given by

$$\tilde{P}_{k_h}[\theta(\tau)] = \sum_{s\epsilon k_h} w(\hat{\tau}_s)\,\tilde{\phi}(\tau|\hat{\tau}_s)\left[\sum_{s=1}^{N} w(\hat{\tau}_s)\,\tilde{\phi}(\tau|\hat{\tau}_s)\right]^{-1}, \qquad [49]$$

where $\phi(\tau|\hat{\tau}_s)$ is the estimated conditional density function of τ, given $\hat{\tau} = \hat{\tau}_s$.

In a special case where $w(\hat{\tau}_s) = 1$ for all $\hat{\tau}_s$'s, Equation 49 can be reduced to

$$\tilde{P}_{k_h}[\theta(\tau)] = \sum_{s\varepsilon k_h} \tilde{\phi}(\tau|\hat{\tau}_s) \left[\sum_{s=1}^{N} \tilde{\phi}(\tau|\hat{\tau}_s)\right]^{-1} . \qquad [50]$$

The procedure which adopts Equation 50 instead of Equaton 49 is called the Simple Sum Procedure.

Let $p(s\varepsilon k_h)$ be the probability with which the examinee s belongs to the item response subgroup k_h. This can be estimated by the proportion of examinees who belong to a specified discrete item response k_h. In the Proportioned Sum Procedure, for the estimated operating characteristic of the discrete item response k_h to the unknown test item h

$$\tilde{P}_{k_h}[\theta(\tau)] = \sum_{s\varepsilon k_h} \tilde{p}(s\varepsilon k_h)\, \tilde{\phi}(\tau|\hat{\tau}_s) \left[\sum_{s=1}^{N} \tilde{\phi}(\tau|\hat{\tau}_s)\right]^{-1} \qquad [51]$$

can be written, where $\tilde{p}(s\varepsilon k_h)$ is the proportion thus obtained as the estimate of the probability $p(s\varepsilon k_h)$.

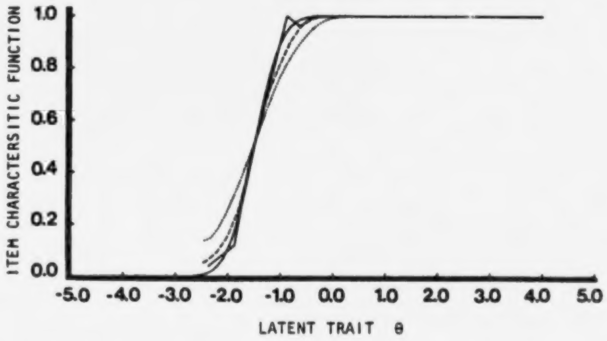
The conditional density, $\phi(\tau|\hat{\tau}_s)$, is approximated by the Pearson System Method, the Normal Approach Method, or the Two-Parameter Beta Method. In so doing, the conditional moments are obtained by Equations 42 through 45, with the replacement of the density function, $g_{k_h}(\hat{\tau})$, by $g(\hat{\tau})$, the density function of $\hat{\tau}$ for the total population of examinees. This is the primary advantage of the Conditional P.D.F. Approach, for the estimation of the marginal density function of $\hat{\tau}$ is done only for the total population, instead of for each item response subpopulation of for each unknown test item.

Figure 3 illustrates three examples obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach, which is combined with the Normal Approach Method. The simulated data used in these examples are the same as those used in the examples given in the preceding section. The results are for the Degree 4 case, and similar results were also obtained for the Degree 3 case.

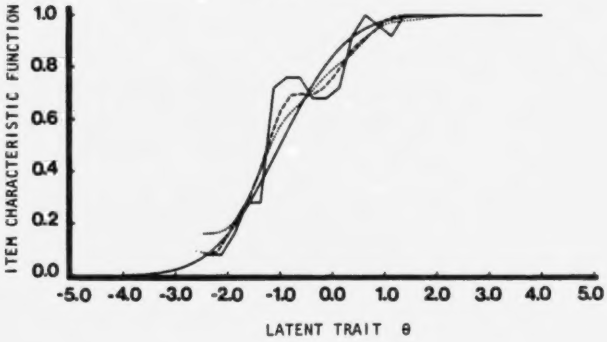
It can be seen, again, that the estimated item characteristic function is fairly close to the theoretical item characteristic function, and much more so in comparison to the corresponding frequency ratios, for each of the three unknown test items. In each of the three graphs, also presented is the resultant estimated item characteristic function of the Degree 4 case obtained by using a subtest of the original Old Test as the Old Test. The subtest consisted of 11 items of the original Old Test. Specifically, out of the 35 items, every 3rd item was chosen, starting with the 3rd and ending with the 33rd item (Samejima & Changas, 1981). Note that in this second situation, the original latent trait θ was transformed to τ, since the amount of test information function is not constant for the interval of θ, in which all the 500 hypothethical examinees were located. The resultant estimated item characteristic functions for the three unknown test items are fairly close to the corresponding theoretical item char-

Figure 3
Estimated Item Characteristic Functions Obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach with the Normal Approach Method, Degree 4 Case, Using Test A (Dashed Curve) and Its Subtest of 11 Test Items (Dotted Curve) as the Old Test, Respectively, in Comparison with the Theoretical Item Charcteristic Function (Smooth Solid Curve) and the Frequency Ratios of the Latent Trait θ (Jagged Solid Curve)

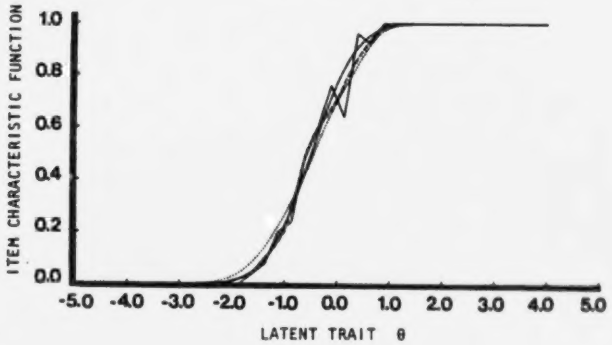
(a) Item 3



(b) Item 4



(c) Item 5



acteristic functions. For Item 4 the fit is even better than the one obtained using the original Old Test, although it is slightly poorer for Item 3 (Figure 3a) and approximately the same for Item 5 (Figure 3c). For each of the other seven unknown test items, the two sets of results turned out to be just as close to each other. That a reasonably good result can be obtained by using only 11 items in the Old Test indicates the robustness of the procedure. Similar studies have been made (Samejima, 1980b, 1981) using a subtest whose test items number 11 and 25. The results turned out to be just as good as those obtained using the original Old Test.

Computer Programs and Model Validation

Seven computer programs have been developed so far for the estimation of the operating characteristics of discrete item responses:

1. TAU TRANSFORMATION
2. MLE THETA
3. CONDITIONAL MOMENTS MLE
4. SIMPLE/WEIGHTED SUM NT
5. PROPORTIONED SUM NT
6. CONDITIONAL MOMENTS SUBGROUP
7. BIVARIATE P.D.F. NT

Figure 4 explains how to combine these programs in order to apply the Conditional P.D.F. Approach of the Bivariate P.D.F. Approach, which are combined either with the Normal Approach Method or with the Two-Parameter Beta Method. Each program is written in such a way that all the input data are printed separately from the results, and the user should pause and examine the result carefully and make some important decisions, if necessary, before proceeding to the next program.

Figure 4
Seven Programs for the Conditional P.D.F. Approach and for the Bivariate P.D.F. Approach, Each of Which Can be Combined Either with the Normal Approach Method or with the Two-Parameter Beta Method

Conditional P.D.F. Approach Combined with the Normal Approach or the Two-Parameter Beta Method:

TAU TRANSFORMATION MLE THETA	} CONDITIONAL MOMENTS MLE {	SIMPLE/WEIGHTED SUM NT PROPORTIONED SUM NT

Bivariate P.D.F. Approach Combined with the Normal Approach or the Two-Parameter Beta Method:

TAU TRANSFORMATION MLE THETA	} CONDITIONAL MOMENTS SUBGROUP	{ BIVARIATE P.D.F. NT

Each program consists of several subroutines. There are three stages for each combination of programs. When the Conditional P.D.F. Approach is used, either SIMPLE/WEIGHTED SUM NT or PROPORTIONED SUM NT is chosen, depending upon whether the Simple Sum or Weighted Sum Procedure, or the Proportion Sum Procedure is used.

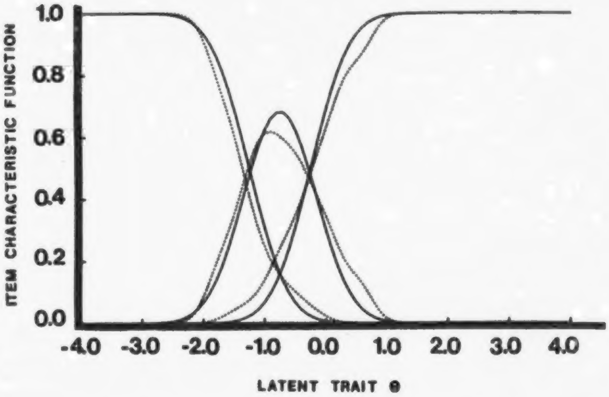
It is noted that the methods and approaches of estimating the operating characteristics of discrete test item responses introduced in the previous sections are directly applicable in such a situation where there is a validated item pool and it is desirable to add more test items to the item pool whose operating characteristics are yet to be discovered. When there is not such an item pool and it is necessary to start from the very beginning of test construction, however, there is no available set of test items which can be used as the Old Test. Thus, it may be necessary to assume some specific model, or models, for the discrete item responses of a set of test items, which were carefully selected with respect to their content validity, and to use them as the Old Test.

Notice that the same methods and approaches that have been developed for estimating the operating characteristics of the discrete item responses of unknown test items can be used for the purpose of model validation for the test items of the Old Test. If the adopted model, or models, is appropriate, then the estimated operating characteristics of the discrete item responses of each item of the Old Test should be close to those which are assumed by the model. This comparison can be performed by deleting a test item, using the remaining $(n - 1)$ test items as the tentative Old Test, and estimating the operating characteristics of the discrete item responses of the deleted test item. If it is found that none of the resultant *n* sets of such estimated operating characteristics validates the adopted model, some other model will have to be used. If it is found that some, but not all, of the resultant *n* sets of estimated operating characteristics invalidate the adopted model, or models, then those test items may be excluded from the original Old Test, and the same process repeated for the remaining set of test items, and so forth. Research is still in progress using the original Old Test of 35 items which follow the normal ogive model with the parameters shown in Table 3.

For the purpose of illustration, Figure 5 presents the estimated operating characteristics of the three graded item responses of Item 15 of the original Old Test, which were obtained by the Simple Sum Procedure of the Conditional P.D.F. Approach combined with the Normal Approach Method, Degree 4 case, using the remaining 34 graded test items as the Old Test. The programs used in this process were Tau TRANSFORMATION, MLE THETA, CONDITIONAL MOMENTS MLE, AND SIMPLE/WEIGHTED SUM NT. It can be seen in this figure that each estimated operating characteristic is reasonably close to the corresponding theoretical operating characteristic.

The simple unweighted least squares method can be used, further, to evaluate the above result, by obtaining the estimates of the two parameters on the normal ogive model on the dichotomous response level, whose item characteristic function is given by Equation 23, as was done previously (Samejima, 1977c, 1977e). In the present example, the estimated operating characteristic, $\tilde{P}_{x_g}(\theta)$,

Figure 5
Estimated Operating Characteristics for the Three Item Score Categories (Dotted Curves) of Item 15 of the Original Old Test, Using the Remaining 34 Graded Test Items as the Old Test in Comparison with the Theoretical Operating Characteristics (Solid Curves) Using the Simple Sum Procedure of the Conditional P.D.F. Approach with the Normal Approach Method



for $x_g = 1$, i.e., the intermediate item score, is added to the one for $x_g = 2$, to produce the estimated item characteristic function, which would be used if it were decided to restore the item dichotomously, assigning the binary item score 1 to $x_g = 1$ and $x_g = 2$ and 0 to $x_g = 0$. In a similar manner, the estimated operating characteristic $\tilde{P}_{x_g}(\theta)$ for $x_g = 2$ (i.e., the highest item score) can be treated as the estimated item characteristic function, which would be adopted if the item were rescored dichotomously, assigning 1 to $x_g = 2$ and 0 to $x_g = 1$ and $x_g = 0$. Thus, there are two estimated item characteristic functions for Item 15, having, presumably, the same discrimination parameter a_g and two different difficulty parameters b_g.

Let $\tilde{P}_g(\theta_j)$ be the estimated value of the item characteristic function in the normal ogive model for the midpoint of the jth interval, and define ζ_{gj} such that

$$\tilde{P}_g(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\zeta_{gj}} e^{-\frac{u^2}{2}} du . \qquad [52]$$

Following the least squares principle, Q is defined by

$$2Q = \sum_{j=1}^{m} [\zeta_{gj} - a_g(\theta_j - b_g)]^2 . \qquad [53]$$

Differentiating Q with respect to a_g and b_g and setting the results equal to zero gives

$$\frac{\partial Q}{\partial a_g} = \sum_{j=1}^{m} [\zeta_{gj} - a_g(\theta_j - b_g)](-\theta_j + b_g) = 0 \qquad [54]$$

and

$$\frac{\partial Q}{\partial b_g} = \sum_{j=1}^{m} [\zeta_{gj} - a_g(\theta_j - b_g)]\, a_g = 0 \; . \qquad [55]$$

The above equations provide the estimates of a_g and b_g such that

$$\hat{a}_g = \mathrm{Cov}(\zeta_{gj}, \theta_j)[\mathrm{Var}(\theta_j)]^{-1} \qquad [56]$$

and

$$\hat{b}_g = m_\theta - [\mathrm{Cov}(\zeta_{gj}, \theta_j)]^{-1}\mathrm{Var}(\theta_j)\bar{\zeta}_g \; , \qquad [57]$$

where m_θ and $\bar{\zeta}_g$ are the means of θ_j and ζ_{gj}, respectively. These estimates, $\hat{a}_g$ and $\hat{b}_g$, are to be compared with their respective parameters, a_g and b_g. The estimates of the two item parameters, $\hat{a}_g$ and $\hat{b}_g$, which are given by Equations 56 and 57, were obtained on each of the two estimated item characteristic functions. It should be noted that the estimated item difficulty parameter, $\hat{b}_g$, based on the first estimated item characteristic function is the estimate of b_{x_g} for $x_g = 1$, and that on the second estimated item characteristic function is the estimate of b_{x_g} for $x_g = 2$.

The estimation was made using the values of the estimated item characteristic function which are less than .9 and greater than .1, obtained for discrete values of θ with the step width of .1. In the first estimation, $\underline{m} = 15$ for the interval of θ, [-1.9, -0.5], and, in the second estimation, $\underline{m} = 18$ for the interval [-1.0, 0.7]. The two estimates of the discrimination parameter, a_g, were 1.790 and 1.421, respectively; these are somewhat less than the true parameter value 2.0 shown in Table 3, the result which is anticipated from Figure 5. The estimates of the two difficulty parameters, b_{x_g} for $x_g = 1$ and $x_g = 2$, were -1.302 and -0.230, respectively. It can be seen that they are reasonably close to the true parameter values, -1.250 and -0.250, given in Table 3.

Thus far the model used for the items of the Old Test has been restricted to the normal ogive model on the graded response level or on the dichotomous response level. Researchers have requested the development of methods of estimating the operating characteristics of discrete item responses using the 3-parameter logistic model (Birnbaum, 1968) for the items of the Old Test or the set of test items whose operating characteristics are known. The item characteristic function, $P_g(\theta)$, of item g, which follows the 3-parameter logistic model is given by

$$P_g(\theta) = c_g + (1-c_g)\Psi_g(\theta) \ , \qquad [58]$$

where $\Psi_g(\theta)$ is the item characteristic function in the logistic model, for which

$$\Psi_g(\theta) = [1 + \exp\{-Da_g(\theta - b_g)\}]^{-1} \ , \qquad [59]$$

can be written and c_g is a constant that is supposed to be the probability of guessing correctly, or unity divided by the number of the alternatives attached to the multiple-choice item g.

There are several theoretical problems concerning the 3-parameter logistic model, including the fact that it does not assure the existence of the unique maximum likelihood estimate of the latent trait θ for every response pattern (cf. Samejima, 1972, 1973b). This is caused by the fact that the item response information function, $I_{k_g}(\theta)$, which is defined by Equation 5 for the discrete item response k_g, assumes negative values for the positive response to the binary item g for the interval of θ less than the critical value θ_g (cf. Samejima, 1973b). This critical value is considerably high, relative to the difficulty parameter b_g. For example, when $D = 1.7$, $a_g = 1.0$, $b_g = 0.0$, and $c_g = .20$, the critical value θ_g is as high as .473. If the value of c_g is changed from .20 to .25, i.e., the item is changed from a five-alternative multiple-choice test item to a four-alternative one, this gives $\theta_g \doteq -0.408$.

These facts indicate that a test item which follows the 3-parameter logistic model has a serious problem caused by the random guessing noise. In fact, in the above example, $P_g(\theta) = 0.5$ at $\theta = \theta_g$, which can be partitioned into two parts, i.e., .333 for knowledge and .167 for guessing correctly. This indicates that the effect of noise is as large as 50% of that of the true information. The use of the test items following the 3-parameter logistic model for the Old Test has this serious problem, and methods should be developed in such a way that test items of lower levels relative to the individual examinee's ability level must be excluded from the maximum likelihood estimation of his or her ability. It can be forseen that the method will necessarily be much more complicated than those in which a model such as the normal ogive model or the logistic model is used.

Information Loss Caused by Asymptotes

It has been seen in the preceding section that the 3-parameter logistic model, in which the item characteristic function has a lower asymptote greater than zero, has the interval, $(-\infty, \theta_g)$, for which the item response information function for the positive response assumes negative values. Even in such a situation, the item information function, $I_g(\theta)$, is nonnegative throughout the entire range of θ, as is obvious from Equation 6 (see Table 2). It is questionable, however, if the item information function is just as meaningful in such a situation as it is when all the item response information functions are nonnegative.

For the moment, the above question will be kept open, and investigation will continue into the amount of the loss of information caused by the lower asymptote which is greater than zero, or by the upper asymptote which is less than unity. In so doing, there will only be consideration of item characteristic functions that are strictly increasing in θ, with c_{g1} and c_{g2} as the lower and upper asymptotes, respectively. Thus,

$$\left\{\begin{array}{l} \lim\limits_{\theta\to\underline{\theta}} P_g(\theta) = c_{g1} \ , \\ \lim\limits_{\theta\to\bar{\theta}} P_g(\theta) = c_{g2} \end{array}\right. \qquad [60]$$

where $\underline{\theta}$ and $\bar{\theta}$ are the lower and upper endpoints of the interval for which θ is defined, and

$$0 \le c_{g1} < c_{g2} \le 1 \ . \qquad [61]$$

Let $\underline{j}$ be a binary test item with $c_{j1} = 0$ and $c_{j2} = 1$. Define τ such that

$$\tau = \tau(\theta) = P_j^{-1}[P_g(\theta)] \ . \qquad [62]$$

Thus, τ is a strictly increasing function of θ with the range

$$\underline{\tau} < \tau < \bar{\tau} \ , \qquad [63]$$

where

$$\left\{\begin{array}{l} \underline{\tau} = P_j^{-1}[P_g(\underline{\theta})] = P_j^{-1}(c_{g1}) \\ \bar{\tau} = P_j^{-1}[P_g(\bar{\theta})] = P_j^{-1}(c_{g2}) \ . \end{array}\right. \qquad [64]$$

For the two item information functions, $I_g(\theta)$ and $I_j^*(\tau)$,

$$\int_{\underline{\theta}}^{\bar{\theta}} [I_g(\theta)]^{1/2} \, d\theta = \int_{\underline{\tau}}^{\bar{\tau}} [I_j^*(\tau)]^{1/2} \, d\tau \ . \qquad [65]$$

can be written. For convenience, and without loss of generality, let $\underline{j}$ follow the logistic model, such that

$$P_j(\tau) = [1 + \exp\{-Da_j(\tau - b_j)\}]^{-1} \ . \qquad [66]$$

Thus,

$$\begin{cases} \underline{\tau} = [Da_j]^{-1}[\log c_{g1} - \log(1-c_{g1})] + b_j \\ \bar{\tau} = [Da_j]^{-1}[\log c_{g2} - \log(1-c_{g2})] + b_j \ . \end{cases} \qquad [67]$$

From Equation 67 it can be seen that if $c_{g1} = 0$, then $\underline{\tau} = -\infty$, otherwise, $\underline{\tau}$ is finite; and if $c_{g2} = 1$, then $\bar{\tau} = \infty$, otherwise, $\bar{\tau}$ is finite. From Equation 65 and 66 for the total information Q is obtained:

$$Q = \int_{\underline{\theta}}^{\bar{\theta}} [I_g(\theta)]^{1/2} \, d\theta$$

$$= Da_j \int_{\underline{\tau}}^{\bar{\tau}} [\exp\{Da_j(\tau-b_j)\}]^{1/2} [1 + \exp\{Da_j(\tau-b_j)\}]^{-1} \, d\tau \ . \qquad [68]$$

Now τ^* will be further defined such that

$$\tau^* = \tau^*(\tau) = [\exp\{Da_j(\tau-b_j)\}]^{1/2} \ . \qquad [69]$$

Thus,

$$\frac{d\tau}{d\tau^*} = 2(Da_j\tau^*)^{-1} \ , \qquad [70]$$

and from Equations 67 and 69

$$\underline{\tau}^* = \tau^*(\underline{\tau}) = c_{g1}^{\ 1/2} (1-c_{g1})^{-1/2} \qquad [71]$$

and

$$\bar{\tau}^* = \tau^*(\bar{\tau}) = c_{g2}^{\ 1/2} (1-c_{g2})^{-1/2} \ . \qquad [72]$$

can be written. From Equations 68, 69, 70, 71, and 72 is obtained

$$Q = 2[\tan^{-1}\{c_{g2}/(1-c_{g2})\}^{1/2} - \tan^{-1}\{c_{g1}/(1-c_{g1})\}^{1/2}] \ . \qquad [73]$$

It is obvious from Equation 73 that when $c_{g1} = 0$ and $c_{g2} = 1$, as is the case with the normal ogive model or with the logistic model on the dichotomous response level, this quantity, Q, assumes the maximum value, such that

$$Q = \pi \ . \qquad [74]$$

It can also be seen from Equation 73 that as c_{g1} departs from zero and c_{g2} from unity, the total information Q becomes progressively smaller than π.

When item g follows the 3-parameter logistic model, or 3-parameter normal ogive model, this gives

$$\begin{cases} c_{g1} = c_g > 0 \\ c_{g2} = 1 \ . \end{cases} \qquad [75]$$

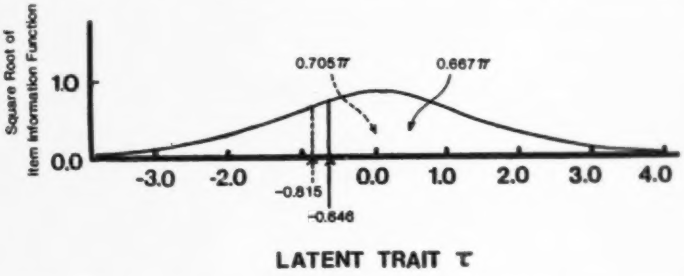
From Equation 73 and 74 can be written

$$Q = \pi - 2\tan^{-1}[c_g/(1-c_g)]^{1/2} \ . \qquad [76]$$

Thus, it can be seen that the second term in the right-hand side of Equation 76 indicates the information loss caused by the asymptote c_g, which is greater than zero. This information loss is as large as $.295\pi$ for $c_g = .20$ (i.e., when the multiple-choice test item has five alternative answers); when $c_g = .25$ (i.e., the multiple-choice test item has four alternative answers), it is as large as $.333\pi$, which indicates that one-third of the total information is lost.

Figure 6 illustrates the information loss caused by the lower asymptote, which is greater than zero. In this example, the solid curve indicates the square root of the item information function in the logistic model, with $D = 1.7$, $a_j = 1.0$, and $b_j = 0.0$. The total area under the solid curve above the abscissa equals π, the left-hand side of the vertical dashes indicates the information loss caused by $c_j = .20$, and the left-hand side of the vertical solid line corresponds to the information loss caused by $c_j = .25$.

Figure 6
Information Loss Caused by the Lower Asymptote c_g of the Item Characteristic Function in the 3-Parameter Logistic Model for $c_g = .20$ (Dashes) and for $c_g = 0.25$ (Solid Line)



The information loss caused by the lower asymptote c_g was computed for each of the three additional values of c_g, i.e., .10, .33, and .50; they are all presented in Table 5 together with the corresponding values of $\underline{\tau}$. In the same table, also presented are the critical value θ_g; the values of $\Psi_g(\theta)$ and $P_g(\theta)$; and that of the item information function, $I_g(\theta)$, at $\theta = \theta_g$, for which

$$\theta_g = (2Da_g)^{-1} \log c_g + b_g \ , \qquad [77]$$

$$\Psi_g(\theta_g) = [1 + \exp\{-Da_g(\theta_g - b_g)\}]^{-1} = c_g^{1/2} (1 + c_g^{1/2})^{-1} \ , \qquad [78]$$

$$P_g(\theta_g) = c_g + (1-c_g)\Psi_g(\theta_g) = c_g^{1/2} \ . \qquad [79]$$

and

$$I_g(\theta_g) = (1-c_g)D^2 a_g^2 \Psi_g(\theta_g)^2 [1-\Psi_g(\theta_g)][c_g + (1-c_g)\Psi_g(\theta_g)]^{-1}$$

$$= D^2 a_g^2 c_g^{1/2} (1-c_g^{1/2})(1+c_g^{1/2})^{-2} \ . \qquad [80]$$

can be written.

Table 5
Information Loss Caused by the Lower Asymptote of the Item Characteristic Function in the 3-Parameter Logistic Model for Each of the Five Values of the Asymptote c_g Together with $\underline{\tau}$, the Critical Value θ_g, and the Logistic Function $\Psi_g(\theta)$, the Item Characteristic Function $P_g(\theta)$ and the Item Information Function $I_g(\theta)$ at $\theta = \theta_g$

c_g	$\underline{\tau}$	Information Loss	θ_g	$\Psi_g(\theta_g)$	$P_g(\theta_g)$	$I_g(\theta_g)$
.10	-1.292	.644 (20.5%)	.677	.240	.316	.361
.20	.815	.927 (29.5%)	.473	.309	.447	.341
.25	.646	1.047 (33.3%)	.408	.333	.500	.321
.33	.408	1.231 (39.2%)	.323	.366	.577	.283
.50	.000	1.571 (50.0%)	.204	.414	.707	.205

Shiba's Word/Phrase Comprehension Tests and Scale Construction

The 13 word/phrase comprehension tests developed by Shiba (1978), which were introduced by Samejima (1980d), are based upon an item pool of 480 test items. Out of those 13 tests, tests AP1 and AP2 are for young children who have not yet learned how to read, and pictures are used instead of words and phrases. Tests A1, A2, A3, A4, A5, nd A6 are basicially for elementary school children of six different grades. Two tests, J1 and J2, are basically for junior high school students of three different grades; and Tests S1 and S2 are for senior high school students of three different grades. Finally, Test U2 is for college students and adults. Each test item is a multiple-choice item, with five alternative answers, i.e., one correct answer and four distractors. The distractors were carefully chosen in such a way that the choice of each alternative will provide information, as does the choice of the correct answer.

Shiba's item pool of 480 items is constructed so that a certain number of test items are used in two adjacent tests; Table 6 presents the total number of test items in each test, together with the number of test items which are shared by each pair of adjacent tests. As can be seen, the numbers of test items for separate tests vary between 30 and 60, and the numbers of shared test items are between 8 and 20.

Table 6
Number of Items of Each of Shiba's 13 Tests of Words/Phrase Comprehension, and Number of Shared Items for Each Pair of Tests

Test	Number of Items	No. of Shared Items
AP1	30	
		10
AP2	30	
		8
A1	36	
		8
A2	34	
		16
A3	40	
		8
A4	40	
		16
A5	48	
		16
A6	54	
		16
J1	56	
		20
J2	60	
		20
S1	60	
		20
S2	60	
		20
U2	57	

When an item pool of a broad range of difficulty levels, such as Shiba's, is available, a number of different uses can be made of it, which may lead to fruitful research results. In so doing, scale construction with a sound theoretical background is most desirable. If this turns out to be successful, then it will be possible to compare, for instance, the ability of two individuals who differ substantially in age and education, to conduct developmental research by following up single individuals, and so forth.

As the first trial, Shiba and others factor analyzed each of 11 of the tests (excluding AP1 and AP2) separately, using a set of data obtained upon a single group of examinees, and then combined the results by equating the item parameters of overlapping test items. The normal ogive model, or the logistic model as its approximation, was assumed for each multiple-choice test item which was scored dichotomously. They were aware that the general factor, which they found for each set of a test and a group of examinees, might not be the same general factor they found for another set, and it is an assumption that all those general factors are one single factor.

To ameliorate this relative weakness, it is conceivable to apply simultaneous factor analysis for each subset of test items which is shared by two adjacent tests, using two or more different groups of examinees. If it turns out that a general factor exists across two or more groups of examinees, then it will be stronger support for the unidimensionality of the latent trait.

In addition to the above benefit, there is a strong possibility that a subset of overlapping test items can be used as the Old Test, from which can be estimated the operating characteristics of both the correct answers and distractors of the other nonoverlapping test items of the two adjacent tests (Shiba, Noguchi, & Haebara, 1978). If the resultant estimated operating characteristics turn out to be similar to the plausibility functions provided by the family of models developed for the multiple-choice item (Samejima, 1979b), then use can be made of the information given by the distractors, as well as the information provided by the correct answer, in estimating the examinee's latent trait. The same process can be also applied for estimating the operating characteristics of the distractors of the test items of the Old Test itself, as well as the model validation for their correct answers, in a similar manner which was illustrated in the sixth section.

The estimation of the operating characteristics will be more accurate if a test has two subsets of shared test items, as is the case with most of Shiba's tests, on lower and upper sides of the latent trait. Through the results of this estimation, it will be possible to equate the scales more rigorously, not only through the sets of shared subsets of test items but through all the other test items.

In applying simultaneous factor analysis, it is necessary that there are covariance matrices, instead of correlation matrices, for what is needed is a common scale for each item variable across the two or more populations of examinees involved. In a situation such as this, however, the covariance matrices are not readily available, since the item variable, which is denoted by X_g, is a hypothesized variable behind the item score. On the other hand, with a normal distribution assumption, the tetrachoric correlation matrix can be obtained for each group of examinees. A way must be found, therefore, to convert tetrachoric correlation matrices for two or more populations to covariance matrices.

There is a simple, straightforward solution for the above problem when a distractor or distractors deserving the second best item score can be located. Here, a simple case will be considered in which there are only two populations of examinees, to whom a common subtest, including item g, was administered. It is assumed that the item variable, X_g, distributes normally for each of the two populations of examinees, with μ_{g1} and σ_{g1} as the two parameters for Populations 1 and μ_{g2} and σ_{g2} for Populations 2. Let p_{gr1} and p_{gr2} be the probabilities with which the examinee chooses the alternative with the rth highest item score for item g, for Populations 1 and 2, respectively, and γ_{gr1} and γ_{gr2} denote the corresponding normal deviates, such that

$$\begin{cases} \gamma_{gr1} = \Phi^{-1}[1 - \sum_{s=1}^{r} p_{gs1}] \\ \gamma_{gr2} = \Phi^{-1}[1 - \sum_{s=1}^{r} p_{gs2}] \ , \end{cases} \qquad [81]$$

where Φ indicates the standard normal distribution function. Thus, for the ratio of the two standard deviations

$$\sigma_{g2}/\sigma_{g1} = (\gamma_{gr1} - \gamma_{gs1})/(\gamma_{gr2} - \gamma_{gs2}) \ . \qquad [82]$$

can be written. When one set of distractors which deserves the second best item score can be located, r = 2 can be set for the correct answer and s = 1 for the distractors, and an estimate of the ratio of the standard deviations of the item variable X_g can be obtained for Populations 1 and 2 through Equation 81, by using the corresponding sample proportions for p_{gr1}, p_{gr2}, p_{gs1}, and p_{gs2}. When, in addition, another set of distractors which deserves the third best item score, and so on, can be located, there is more than one set of relationships described by Equation 81. In such a case, some least square procedure will provide an estimate of the ratio of the two standard deviations. In so doing, it is desirable to allow errors on both sides such as is done by Deming's (1946) least squares method. For the distance between the two means of X_g

$$\mu_{g2} - \mu_{g1} = \sigma_{g1}\gamma_{gr1} - \sigma_{g2}\gamma_{gr2} \ . \qquad [83]$$

can be written. Since there is more than one value of r in Equation 83, the arithmetic mean of the results of Equation 83 will provide an estimate of the distance between the two means.

Table 7 presents two contingency tables for the choice of the alternatives of Item 51 of Shiba's Test J1 against five test score groups, for the samples of seventh graders and eighth graders, respectively. It can be seen in these two tables that, next to the correct answer E, Alternative C attracted examinees of

Table 7
Contingency Tables of the Five Alternatives--A, B, C, D, and E--of Item 51 of Test J1 vs. Five Test Score Categories Together with the Mean of Maximum Likelihood Estimate of θ for Each Alternative Subgroup for Seventh- and Eighth-Grade Students

Subgroup and Test Score	Alternative A	B	C	D	E*	No Answer	Total
Seventh Grade							
Lowest	27	36	11	30	22	4	130
Low	10	20	12	22	0	3	107
Middle	23	10	25	28	32	2	120
High	21	16	18	24	33	6	118
Highest	22	4	12	21	79	1	139
Total	103	86	78	125	206	16	614
Mean $\hat{\theta}$	.137	.640	.077	.136	.429	—	
Eighth Grade							
Lowest	25	18	10	17	12	3	85
Low	13	14	9	16	27	2	81
Middle	17	8	14	16	48	5	108
High	10	6	18	13	38	2	87
Highest	11	1	4	6	77	1	100
Total	76	47	55	68	202	13	461
Mean $\hat{\theta}$	.118	.260	.312	.150	.909	—	

*Correct Answer

Figure 7
Relationship among the Proportions of Examinees Who Correctly Answered Item 4 of Test J1 and Who Selected the Second Best Alternative, and the Means and Standard Deviations of the Two Populations JH1 and JH2

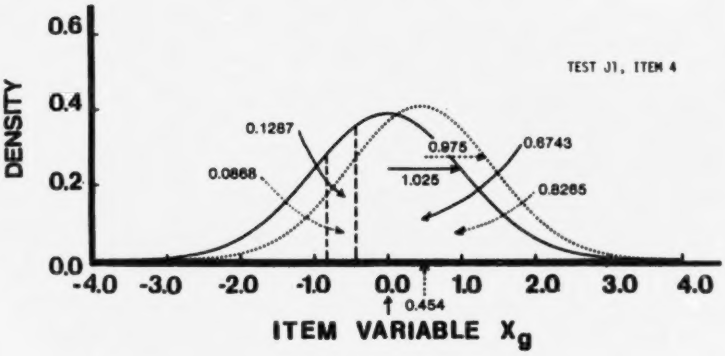


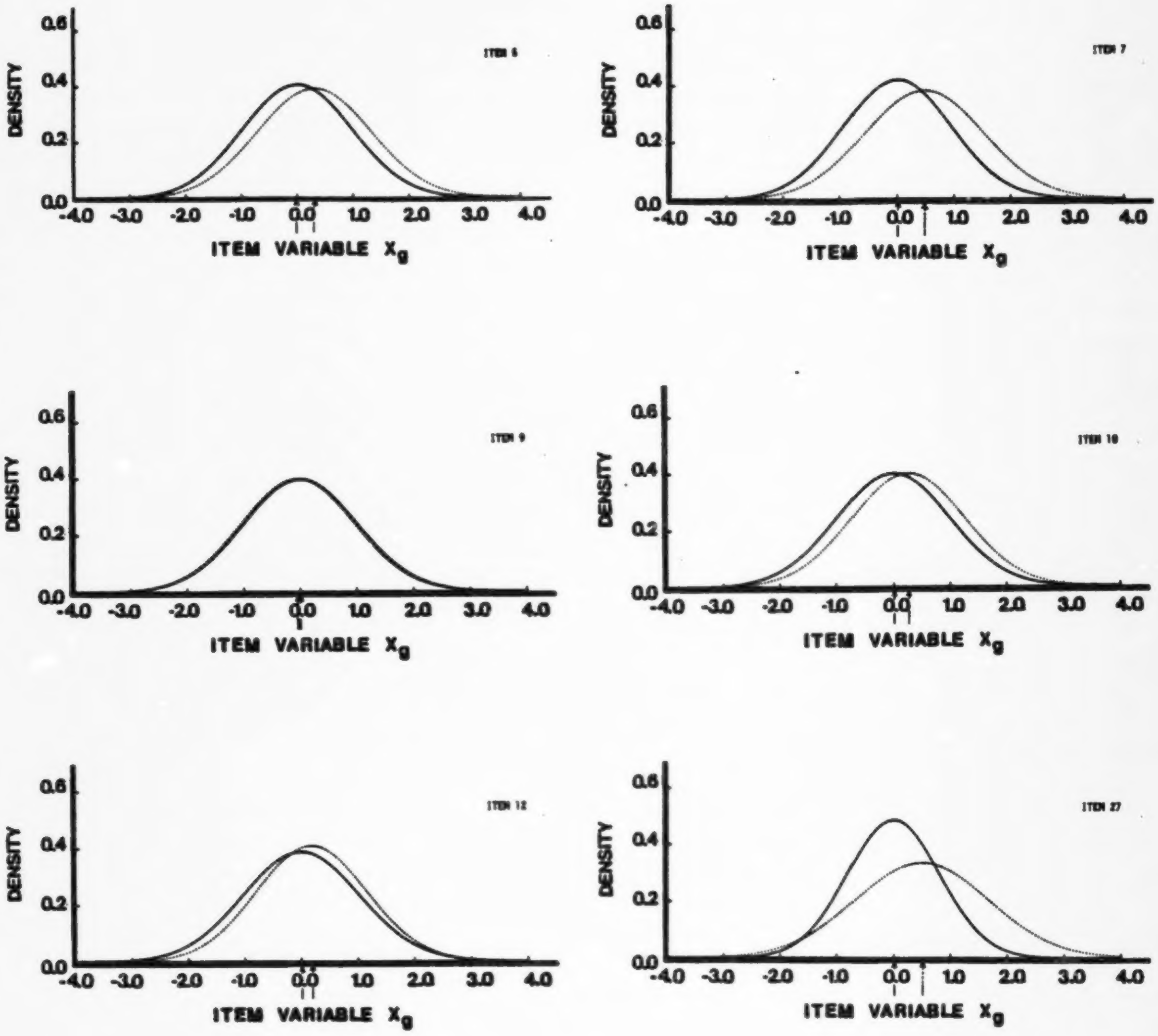
Table 8
Estimated Means and Standard Deviations of Each of the 14 Item Variables of Test J1 for the Two Populations, JH1 and JH2, Together with the Correct Answer and Alternative(s) of the Second Best Item Score

	Mean		SD		Correct	Second
Item	JH1	JH2	JH1	JH2	Answer	Best
4	.000	.454	1.025	.975	E	A
5	.000	.308	.982	1.018	A	D, E
7	.000	.479	.953	1.047	D	B
9	.000	-.047	1.002	.998	C	D, E
10	.000	.266	1.000	1.000	E	B
12	.000	.185	1.026	.974	B	A, E
27	.000	.499	.816	1.184	E	C
29	.000	.131	1.017	.983	C	D
30	.000	.130	.973	1.027	B	A, D
34	.000	.162	.945	1.055	B	D
47	.000	.743	.730	1.270	B	C, E
48	.000	.131	1.008	.992	E	A, B
50	.000	.106	1.037	.963	A	B, D
51	.000	.241	.952	1.048	E	C

Table 9
Factor Loadings of First Three Common Factors Obtained by Simultaneous Factor Analysis of 14 Items of Test J1 Using JH1 and JH2 as Two Populations

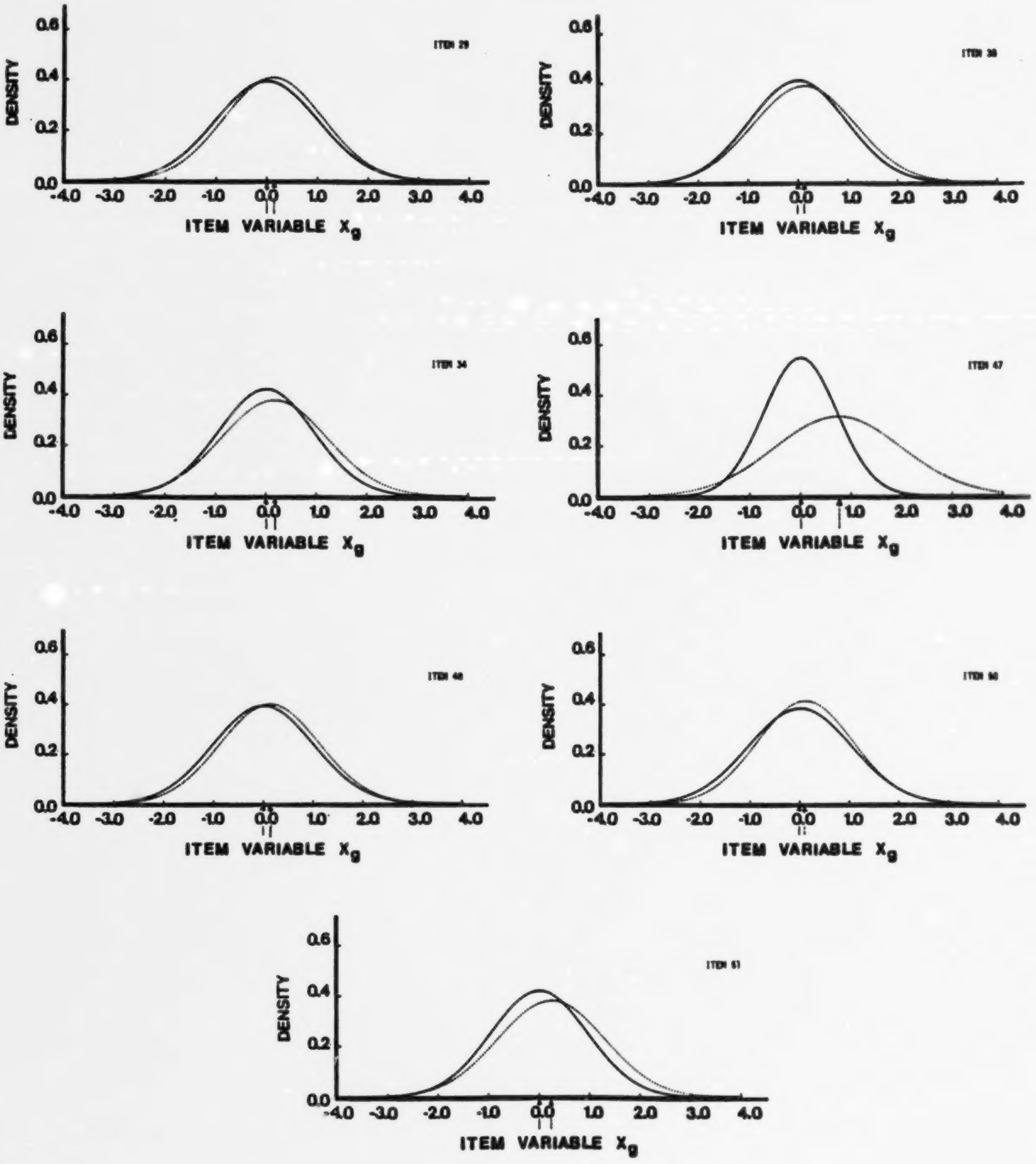
Item	Common Factor		
No.	1	2	3
4	.615	.099	.175
5	.597	-.056	-.178
7	.405	-.330	-.161
9	.340	-.069	-.040
10	.480	-.154	-.418
12	.446	.069	.147
27	.320	-.484	.127
29	.613	-.016	-.271
30	.308	-.070	-.151
34	.286	.048	.066
47	.332	.189	.073
48	.412	-.182	.124
50	.366	-.278	.037
51	.389	-.154	.350

Figure 8
Estimated Density Functions of Item Variable X_g for
the Seventh (Solid Curve) and Eighth (Dotted Curve) Graders
for Each of 13 Items of Test J1



- continued on the next page -

Figure 8 (Continued)
Estimated Density Functions of Item Variable X_g for
the Seventh (Solid Curve) and Eighth (Dotted Curve) Graders
for Each of 13 Items of Test J1



relatively high levels of ability, and Alternative D probably follows C in that tendency. This observation from the contingency tables is supported by each of the two configurations of the means of maximum likelihood estimates θ for the separate alternative subgroups of examinees (cf. Samejima, 1980d), which are given at the bottom of each contingency table. Thus either Alternative C alone, or the combination of Alternative C and D, should be treated as the one which deserves the second best item score.

Figure 7 presents the relationship between the two sets of parameters of the item variable distributions for Item 4 of Test J1, by setting $\mu_{g1} = 0$ and $\sigma_{g1} + \sigma_{g2} = 2$. Fourteen test items have been chosen out of the 56 of Test J1, for each of which the identification of the second best alternative answer, or answers, was made relatively easily. The estimated means and standard deviations for the two populations, JH1 and JH2, for each of the 14 test items, together with the correct answer and the alternative(s) with the second highest item score, are given in Table 8. Figure 8 also presents the estimated density functions of the item variable X_g for the two populations for each of the remaining 13 test items of Test J1, excluding Item 4, which was already illustrated in Figure 7. It can be seen from the table and figures that, except for a few items, such as Items 27 and 47, the two estimated density functions are close to each other.

By multiplying the estimated tetrachoric correlation coefficient between the two item variables, X_g and X_j, with the product of the corresponding two estimated standard deviations, the correlation coefficient can be converted to the covariance of each pair of test items, g and j, and hence the sample correlation matrix to the sample covarince matrix for each population.

Table 10
Covariance Matrices of
the Three Common Factors
Resulting from Simultaneous
Factor Analysis for
Populations JH1 and JH2

Population and Common Factor	Common Factor 1	2	3
JH1			
1	1.000		
2	0.000	1.000	
3	0.000	0.000	1.000
JH2			
1	1.072		
2	-.324	-.091	
3	.354	.393	1.037

The factor loadings of the first three common factors for the 14 test items obtained by the simultaneous factor analysis (SIFASP) for the two examinee groups, JH1 and JH2, are presented in Table 9. They are orthogonal and standardized factors for JH1 and oblique for JH2. The covariance of these three factors for JH1 and JH2 are given as Table 10. It can be seen from Table 9 that the first common factor can be considered as the general factor for these 14 item variables.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Deming, W. E. Statistical adjustment of data. New York: Wiley, 1946.

Elderton, W. P., & Johnson, N. L. Systems of frequency curves. Cambridge England: Cambridge University Press, 1969.

Kendall, M. G., & Stuart, A. The advanced theory of statistics (Vol. 2). New York: Hafner, 1961.

Lawley, D. N., & Maxwell, A. E. Factor analysis as a statistical method. London: Butterworth, 1971.

Levine, M. Appropriateness measurement and the formula-score method: Overview, intercorrelations and interpretations. Paper presented at the Office of Naval Research conference on model-based psychological measurement, Iowa City IA, 1980.

Lord, F. M. A theory of test scores. Psychometric Monograph, No. 7, 1952.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--A confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Rasch, G. Probabilistic models for some intelligence and attainment tests (expanded edition). Chicago: University of Chicago Press, 1980. (Originally published, Copenhagen: Danmarks Paedagogiske Institut, 1960)

Samejima, F. Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17, 1969.

Samejima, F. A general model for free-response data. *Psychometrika Monograph*, No. 18, 1972.

Samejima, F. Homogeneous case of the continuous response level. *Psychometrika*, 1973, *38*, 203-219. (a)

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, 1973, *38*, 221-233. (b)

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 1974, *39*, 111-121. (a)

Samejima, F. *Normal ogive model on the graded response level in the multidimensional latent space.* Paper presented at the annual meeting of the Psycometric Society, Stanford CA, April 1974. (b)

Samejima, F. Graded response model of the latent trait theory and tailored testing. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U.S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Samejima, F. Effects of individual optimization in setting the boundaries of dichotomous items on accuracy of estimation. *Applied Psychological Measurement*, 1977, *1*, 77-94. (a)

Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1977, *1*, 233-247. (b)

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, 1977, *42*, 163-191. (c)

Samejima, F. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika*, 1977, *42*, 193-198. (d)

Samejima, F. *Estimation of the operating characteristics of item response categories I: Introduction to the two-parameter beta method* (Research Report 77-1). Knoxville: University of Tennessee, Department of Psychology, 1977. (e)

Samejima, F. *Estimation of the operating characteristics of item response categories II: Further development of the two-parameter beta method* (Research Report 78-1). Knoxville: University of Tennessee, Department of Psychology, December 1978. (a)

Samejima, F. *Estimation of the operating characteristics of item response categories III: The normal approach method and the Pearson system method* (Research Report 78-2). Knoxville: University of Tennessee, Department of Psychology, June 1978. (b)

Samejima, F. *Estimation of the operating characteristics of item response categories IV: Comparison of the different methods* (Research Report 78-3). Knoxville: University of Tennessee, Department of Psychology, June 1978. (c)

Samejima, F. *Estimation of the oeprating characteristics of item response categories V: Weighted sum procedure in the conditional P.D.F. approach* (Research Report 78-4). Knoxville: University of Tennessee, 1978. (d)

Samejima, F. *Estimation of the operating characteristics of item response categories VI: Proportioned sum procedure in the conditional P.D.F. approach* (Research Report 78-5). Knoxville: University of Tennessee, Department of Psychology, 1978. (e)

Samejima, F. *Estimation of the operating characteristics of item response categories VII: Bivariate P. D. F. approach with normal approach method* (Research Report 78-6). Knoxville: University of Tennessee, Department of Psychology, December 1978. (f)

Samejima, F. *Constant information model: A new, promising item characteristic function* (Research Report 79-1). Knoxville: University of Tennessee, Department of Psychology, February 1979. (a)

Samejima, F. *A new family of models for the multiple-choice item* (Research Report 79-4). Knoxville: University of Tennessee, Department of Psychology, December 1979. (b)

Samejima, F. *Estimation of the operating characteristics when the test information of the old test is not constant I: Rationale* (Research Report 80-2). Knoxville: University of Tennessee, Department of Psychology, June 1980. (a)

Samejima, F. *Estimation of the operating characteristics when the test information of the old test is not constant II: Simple sum procedure of the conditional P. D. F. approach/normal approach method using three subtests of the old test* (Research Report 80-4). Knoxville: University of Tennessee, Department of Psychology, November 1980. (b)

Samejima, F. Latent trait theory and its applications. In P. R. Krishnaiah (Ed.), *Multivariate Analysis V*. Amsterdam, Netherlands: North-Holland, 1980. (Paper presented at the Fifth International Symposium on Multivariate Analysis, University of Pittsburgh, Pittsburgh PA, 1978.) (c)

Samejima, F. *Research on the multiple-choice test item in Japan: Toward the validation of mathematical models* (ONRT-M3). Tokyo: Department of the Navy, Office of Naval Research, April 1980. (d)

Samejima, F. *Estimation of the operating characteristics when the test informa-*

tion of the old test is not constant II: Simple sum procedure of the conditional p.d.f. approach/normal approach method using three subtests of the old test (Research Report 81-2). Knoxville: University of Tennessee, Department of Psychology, July 1981.

Samejima, F., & Changas, P. S. *How small the number of test items can be for the basis of estimating the operating characteristics of the discrete responses to unknown test items* (Research Report 81-3). Knoxville: University of Tennessee, Department of Psychology, November 1981.

Samejima, F., & Livingston, P. *Method of moments as the least squares solution for fitting a polynomial* (Research Report 79-2). Knoxville: University of Tennessee, Department of Psychology, June 1979.

Shiba, S. Construction of a scale for acquisition of word meanings. *Bulletin of the Faculty of Education, University of Tokyo*, 1978, *17*, 47-58. (In Japanese with English abstract)

Shiba, S., Noguchi, H., Haebara, T. A stratified adaptive test of verbal ability. *Japanese Journal of Educational Psychology*, 1978, *26*, 11-20. (In Japanese with English abstract)

Tucker, L. R. *Academic ability test* (Research Memorandum 51-17). Princeton NJ: Educational Testing Service, 1951.

DISCUSSION

RODERICK P. McDONALD
MACQUARIE UNIVERSITY

Anyone who reads Samejima's paper "Development and Application of Methods for Estimating Operating Characteristics of Discrete Test Item Responses without Assuming Any Mathematical Form" in its complete form must be impressed by the depth and carefulness of her treatment of this problem. I see the problem in question perhaps in an unusual fashion, as an analog of factor extension in linear common factor analysis. In factor extension the known factor loadings of a core set of tests are used, and joint covariances of the new set (the extension set) determine the factor loadings of the extension set.

Samejima's problem differs from this in two ways. First, the IRT model is nonlinear. Second, a prescribed mathematical function form is used for the operating characteristic of the core set of items (the Old Test in her terminology) but not for the extension set of items. The regressions of the new items on the θ of the old items are represented graphically or numerically. It is only the second of these differences that seems significant to me.

Speaking out of my personal tastes in theorizing, I see the lack of a mathematical function for the new set as a disadvantage to the user. It is balanced by the advantage that motivates the work, that is, that the extension set can contain any kind of multinomial (multicategory) variables. I would need further reasons for adopting this approach rather than directly fitting suitable regressions for multinomial variables on previously estimated latent trait values, since I would expect to use the item parameters of the extension items in further work and would not know how to use these graphs as such.

McDonald and Ishizuka (unpublished) have proposed a hyperbolic transformation of the logistic function that provides a general set of multinomial item characteristic curves. Essentially, the hyperbolic treatment controls a lower and an upper asymptote for each category and the category probabilities can be arranged to sum to unity everywhere, yielding a slightly more flexible scheme than the one that Bock (1972) described for multinomial variables. We have not worked further on this because the fundamental problem is at the latent trait end. I would argue that it would work very well for extension purposes, where essentially one is simply doing regressions with a multinomial dependent variable.

I would also suggest that in this work, as in the factor analog, it is essential to compute and to examine the item residual covariance matrices of the core items and the extension items, as well as the residual cross-covariance matrix, in order to verify the hypothesis of joint unidimensionality. This is because in factor extension work the analogous hypothesis is very commonly false; whether or not the extension will be co-unidimensional with the original set is a more sensitive question in extension work than in straight factor analysis.

Faced with the same class of problems as tackled here, I would have looked at the frequency ratios that were evident in the jagged graphs of the new items plotted against the old θ. I would have agreed that the graph needs smoothing, and I would have chosen some standard method such as spline functions to obtain a smooth graph. It is possible to conclude that Samejima has developed ideal methods against which the cruder standard devices such as regression splines could be tested and compared.

Reference

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

THE TRAIT IN LATENT TRAIT THEORY

MICHAEL V. LEVINE
UNIVERSITY OF ILLINOIS

Significant to a latent trait or item response theory analysis of a mental test is the determination of exactly what is being quantified. The following are practical problems to be considered in the formulation of good theory:

1. Deciding whether two tests measure the same trait or traits;
2. Analyzing the relative contributions of a pair of traits or abilities to test performance;
3. Detecting "functional" changes in items including those caused by security problems, mode of administration changes, and changes in familiarity with the concepts supporting the item in the population being tested;
4. Determining the adequacy of an item response function, i.e., a specific mathematical formula relating performance to ability;
5. Discovering the shape of the item response functions, including multidimensional item response functions;
6. Quantifying the magnitude and reliability of violations of the principal assumption of latent trait theory, "local independence";
7. Modeling item responses (such as omitting or changing answers) that fail to be locally independent.

In this paper theoretical results bearing on these problems will be outlined. A new theory will be presented, with the central problem the representation of traits, abilities or achievements, and their distributions.

1. Motivation

Three Practical Problems

To motivate the new theory, three important measurement problems will be used. First, however, two latent trait theory terms, "item response function" and "local independence," will be defined for the special situations considered in this section (to be redefined in later sections where more generality is needed).

The item response function (IRF; also called the item characteristic curve and conditional response function) is the (conditional) probability of sampling an examinee correctly answering the item from the subpopulation of all examinees at a particular ability level. Thus, if ability is unidimensional, the IRF for the ith item on a test is the real function P_i, where $P_i(t)$ = the probability of a correct response to item i from an examinee sampled from all those with ability = t. A pair of items, say the ith and jth, are said to be locally independent if they are independent in subpopulations having no variation in ability, i.e., if for all ability levels *t*, e.g., Prob{items *i* and *j* are both correct |ability = t} equals the product of the item response functions $P_i(t)P_j(t)$.

In the maintenance of testing programs that attempt to give more or less the same test, year after year, to a large number of people (e.g., military entrance and placement programs, college entrance exams, graduate and professional school admissions exams, high school and grade school aptitude and achievement tests, and interest measures such as job satisfaction scales used in industrial settings), three difficult measurement problems inevitably arise. These problems are functional item change, IRF adequacy, and local independence failure.

Functional item change. An item may function differently, i.e., have different psychometric properties, in two test administrations. For example, a vocabulary item requiring exposure to political terminology may seem relatively easy in a presidential election year. School curriculum changes, security problems, method of administration change, improper coaching, and item format change also may result in functional item change. The principal question is to determine to what extent, if at all, an item has functionally changed.

IRF adequacy. Many mathematical formulas have been proposed to represent IRFs. Psychological arguments have been used to challenge the correctness of each, usually over an ability range. For example, monotonic curves have been criticized for giving an incorrect representation over a low ability range because very low ability examinees may perform somewhat better than examinees just bright enough to be misled by item construction tricks. Curves that asymptote to 1.0 have been criticized because they contain no provision for the careless mistakes of very bright examinees answering items beneath their ability level. Of course, with the very large samples of examinees, virtually any guess or estimate of the population IRF can be rejected. The goal in adequacy problems is to determine whether a proposed curve is "adequate," i.e., close enough to the population IRF over an ability range to be acceptable in a specific application.

Local independence failure. Psychological reasoning or data analysis can sometimes lead to the suspicion that the local independence assumption has been seriously violated. For example, a pair of reading comprehension items referring to the same reading passage may be, to an unacceptable degree, measuring familiarity with the content of the passage. An example arising in an empirical item bias study is described later in this section. One application of the theory being developed is to determine the magnitude and reliability of suspected local independence failures.

Each of the three problems will be considered separately. Table 1 summarizes the discussion. It will be shown that all the problems can be reduced to a single question about two curves or functions of abilities: How close is a specified (either by a formula or table) function $L(\cdot)$ to an incompletely specified function $P(\cdot)$, the population IRF? Since abilities are only estimated and not observed, the question is difficult to answer.

IRF Adequacy

Consider the conceptually simplest problem type, IRF adequacy. A 3-parameter logistic function L has been estimated and offered as a representation of

Table 1
Summary of Formula Score Theory Analysis
of Three Basic Measurement Problems

Problem	Hypothesis	Population Parameter	Distribution of Test Statistic η
Functional Change	$L^* = P$	$\eta = \int_a^b (P-L)^2$	Quadratic function of normal variables; noncentral case.
Item Response Function Adequacy	$L = P$	$\eta = \int_a^b (P-L)^2$	Quadratic function of normal variables; central case.
Local Independence	$P_1P_2 = P$ $P_1 = L_1$ $P_2 = L_2$	$\eta = \int_a^b (P-L_1L_2)^2$	Quadratic function of normal variables; central case.

the true, i.e., population, IRF P. The psychometrician is concerned about the monotonicity of P and suspects that L fits P poorly over, say, the ability range $-3 \leq \theta \leq -2$. He/she wishes to determine how far apart P and L are over this range.

An intuitive and commonly used measure of the distance between two functions is the generalization of Euclidean distance given by the root mean square of function values. In this spirit, an attempt will be made to compute a point estimate and confidence interval for

$$\eta = \int_{-3}^{-2} [P(t) - L(t)]^2 dt \ . \qquad [1]$$

The interval $[-3,-2]$ in the definition of η is arbitrary. The hypothesis being tested and the sample of examinees available for testing will generally suggest a different center and width of the "supporting" interval. Short intervals give more specific information about the difference between P and L. Very short intervals give estimates of η with a large sampling error.

The results of this section are made possible by an elementary equation which is valid at each point $\underline{t}$ where ability densities are continuous.

$$[P(t) - L(t)]f(t) = f^+(t)[1 - L(t)]\bar{P} + f^-(t)L(t)\bar{Q} \qquad [2]$$

where

P and L are as defined previously;
$\underline{f}$ is the density for the ability θ in the general population of examinees;
f^+ is the ability density in the subpopulation of examinees passing the target item;
f^- is the conditional density in the failure subpopulation; and

$\overline{P} = 1 - \overline{Q}$ is the proportion of item passers.

This equation is important because it permits the evaluation of adequacy questions without estimating abilities. It is necessary to do so because for a test of fixed length, any estimate of ability has a substantial standard error for which a bound can be computed by routine methods. On the other hand, subject to technical qualifications treated at length below, an arbitrarily accurate estimate of the distribution of abilities can be obtained with a sufficiently large sample of examinees, and test administrations of over 1,000,000 examinees are no longer uncommon. In later sections, consistent estimates of ability densities are discussed.

In view of the very large sample sizes, f will be regarded as known. The effects of small errors in specifying f on the sampling distribution of the estimate of η has not been worked out at this time.

The quantity $\overline{P}$ on the right-hand side can either be computed as $\int L(t)f(t)dt$ when the hypothesis P = L is being evaluated or be estimated as the sample proportion of persons who correctly answered the item.

In most applications only moderately large samples are available for estimating the conditional densities f^+ and f^-. Using Equation 2,

$$\eta = \int_a^b \{f^+(t)[1-L(t)]\overline{P} + f^-(t)L(t)\overline{Q}\}^2 W(t)dt \qquad [3]$$

can be written, where the weight function W(t) is $1/[f(t)]^2$. Upon substituting estimates $\hat{f}^+(\cdot)$ and $\hat{f}^-(\cdot)$ for $f^+(\cdot)$ and $f^-(\cdot)$, a statistic $\hat{\eta}$,

$$\hat{\eta} = \int_a^b \{\hat{f}^+(t)[1-L(t)]\overline{P} + \hat{f}^-(t)L(t)\overline{Q}\}^2 W(t)dt \qquad [4]$$

is obtained that can be used to evaluate IRF adequacy. It will be seen that $\hat{\eta}$ has a tractable sampling distribution.

To motivate some theoretical developments on density representation and estimation in the next section, suppose the conditional densities could be represented in the form

$$f^+(\theta) = \sum_{j=1}^{J} \alpha_j^+ h_j(\theta) \qquad [5]$$

and

$$f^-(\theta) = \sum_{j=1}^{J} \alpha_j^+ h_j(\theta) \qquad [6]$$

for known functions $h_1, h_2, \ldots, h_J$ and constants α_j^+, α_j^-. Then, after the indicated integration in Equation 3 is carried out, η has the particularly simple form

$$\eta = \underset{\sim}{\alpha} Q \underset{\sim}{\alpha}^T \qquad [7]$$

where $\underset{\sim}{\alpha}$ is the vector $< \alpha_1^+, \ldots, \alpha_J^+, \alpha_1^-, \ldots, \alpha_J^- >$ and Q is a matrix of numbers that can be calculated prior to data collection. The entries in Q are obtained by substituting Equation 5 and 6 into Equation 3, expanding the product and numerically calculating the integral of the product of the specified functions. It is easily verified that $\underset{\sim}{Q}$ is symmetric, positive, and semidefinite.

Such a representation has been derived. The functions h_j are derived from a priori considerations given in the next section. The number of them, J, turns out to be acceptably small, between 4 and 8, for the tests already analyzed.

Consistent, unbiased estimates for the vector of constants are described in the following sections. With them are obtained estimated densities

$$\hat{f}^+(\cdot) = \sum_{j=1}^{J} \hat{\alpha}_j^+ h_j(\cdot) \qquad [8]$$

and

$$\hat{f}^-(\cdot) = \sum_{j=1}^{J} \hat{\alpha}_j^- h_j(\cdot) \ . \qquad [9]$$

Here $\hat{\alpha}_j^+$ and $\hat{\alpha}_j^-$ are estimates of the corresponding constants.

The vector of estimates $\hat{\underset{\sim}{\alpha}}$ will be seen to be multivariate normal, at least asymptotically. The hypothesis P = L permits calculation, prior to data collection, of the covariance matrix of the estimates and the derivation of the distribution of

$$\hat{\eta} = \hat{\underset{\sim}{\alpha}} Q \hat{\underset{\sim}{\alpha}}^T \qquad [10]$$

Random variables of form $\hat{\underset{\sim}{\alpha}} Q \hat{\underset{\sim}{\alpha}}^T$, where $\hat{\underset{\sim}{\alpha}}$ is multivariate normal and $\underset{\sim}{Q}$ positive, semidefinite, generalize the χ^2 family of random variables. In the "central case" the statistic has the same distribution as the sum of squares of several independent normal variables with zero mean and not necessarily equal variances. In the "noncentral case," the means may be unequal. The asymptotic normality and the hypothesis P = L make the central case appropriate. A numerical algorithm has been developed by the author (with Bruce Williams) for computing the cdf $F(x) = \text{Prob}\{\hat{\eta} < x\}$ and determining the probability of observing an $\hat{\eta}$ equal to the sample value or larger under the hypothesis, $\eta = 0$, i.e., P = L. (For a review of alternative algorithms, see Johnson & Kotz, 1970, chap. 29. Technical details on the derivation and distribution of η are in Levine, 1981.)

Functional Item Change

The above approach can be used to attack functional item change questions. In the treatment of adequacy, the distribution of $\hat{\eta}$ was derived under the hypothesis P = L. In studying change, the discrepancy between P and L is measured under the hypothesis that P = L* where L* is some specified function other than L. Suppose, for example, that an IRF has been carefully measured using a very large sample and that years of sucessful experience with the item were consistent with the IRF used to represent it. The hypothesis P = L* has considerable support. However, after a security problem comes to light, a reestimation with a smaller sample gives a function L ≠ L*. Further, suppose that only low ability examinees are motivated to exploit the security problem. Under the hypothesis P = L*, how large is the squared difference between P and L expected to be over the low ability range? By a generalization of the arguments described above, the distribution of $\hat{\eta}$ can be derived. It turns out to be a quadratic function of normal variables, noncentral case. Formulas for the variances and noncentrality parameters are in Levine (1981).

Local Independence

The method of this section suggests a way to quantify suspected departures from local independence. For example, in an item bias study in progress, a vocabulary item using the word "hurl" was found to be severely and reliably biased against sixth-grade girls and in favor of sixth-grade boys in two independent samples of 4,000 and 2,000 children. It seems likely that performance on another item also using a word favored by baseball writers would agree more with the hurl item score than that predicted by the local independence assumption of latent trait theory. To analyze the causes and consequences of bias, it would be valuable to have a method for measuring the magnitude and reliability of local independence violations over specified ability ranges.

To test for local independence, two suspect items may be (conjunctively) paired to form a complex item--an item that is scored correct if both component items are correct and incorrect otherwise. If the items are locally independent, then the item characteristic curve (ICC) of the complex item will be the product of ICCs of the component items. Evidence for a violation of local independence would be small

$$\eta_1 = \int_a^b [P_1(t) - L_1(t)]^2 dt \qquad [11]$$

and

$$\eta_2 = \int_a^b [P_2(t) - L_2(t)]^2 dt \qquad [12]$$

but large

$$\eta_{12} = \int_a^b [P_{1\&2}(t) - L_1(t)L_2(t)]^2 dt \ . \qquad [13]$$

In all the above examples, densities and conditional densities were represented as linear combinations of a finite set of known functions. It will be shown in later sections that every density is, in a sense soon to be made precise, equivalent to exactly one of these linear combinations. Every test will be shown to have associated with it a unique "canonical space" or vector space of functions equivalent to densities. It is hoped that this discussion shows that a theory for density representation and estimation can be used to attack fundamental issues in psychological measurement.

2. Foundations: Canonical Space and Equivalent Ability Distributions

In the preceding section a relation was noted between several difficult but important substantive psychological issues and the more routine methodological problem of density estimation. The approach of the previous section required a representation of ability densities as finite linear combinations of known functions and multivariate normal estimates of the coefficients in the combinations.

This section informally reviews an a priori derivation of this representation; a more formal presentation of the derivation is outlined in Section 3. This approach to psychometric problems will be called "formula scoring" or "formula score theory" and abbreviated FS and FST.

The analysis is organized about three fundamental theoretical issues and methodological problems:

1. Ability distribution equivalence. Which, if any, pairs of fundamentally different ability distributions lead to exactly the same probability distributions on the item scores--the only observables in testing? What are necessary and sufficient conditions for two distributions to be equivalent (in the sense of making the same predictions)? What statement about the distribution are (in the technical foundations-of-measurement sense of the term) meaningful?
2. Ability distribution representation. Find a decomposition of an arbitrary ability density into two uniquely determined parts

   $$f(\cdot) = f_0(\cdot) + f^*(\cdot) \qquad [14]$$

   such that two densities f_1 and f_2 are equivalent if and only if $f_1^* = f_2^*$. Find a finite dimensional parameterization of the "identifiable part" f^* of the ability density f.
3. Ability distribution identification and estimation. Show that the "identifiable part" of the ability density is identifiable in the sense that a consistent estimate of $f^*(t)$ exists for each t. Construct an estimator.

The results of this section are derived from a version of latent trait theory that is more formally presented in Section 3.

The major random variables of the latent trait model are abilities θ and item scores $u_1, u_2, \ldots, u_n$. Examinees are considered to be randomly sampled from an infinite population of examinees. The "points" in the probability space of the basic latent model are examinees. Each examinee has a specific ability and (nonrandom) vector of item responses. Abilities and item responses are nontrivial random variables only because examinees are sampled. Item responses are assumed to be "locally independent", i.e., independent in the subpopulations of examinees defined by conditioning upon ability. Although it may not be immediately obvious, this conceptualization of latent trait theory is compatible with the usual treatment of item responses as independent binomial random variables with success probabilities that are functionally dependent on abilities, provided no item is ever administered two times to the same examinee.

To attack the problems of identifying, representing, and estimating ability distributions from a foundations-of-measurement point of view, the set of all statistics for a test is studied. Since only the item scores u_i are observed, and since examinees work independently of one another, the set of all statistics is simply the set of number-valued functions of the item score random variables. Moreover, this set can be shown to be a finite dimensional vector space.

An important tool for studying ability distributions in formula score theory is the canonical space of a test, formulated by referring to regression functions. The regression function of a statistic S is the real function

$$R_S(t) = E[S|\theta = t] \ . \qquad [15]$$

The canonical space (CS) of a test is the vector space of all regression functions. It is easily shown to be finite dimensional. In fact, in many FST applications it has been possible to treat it as a vector space of low (less than 8) dimensionality. (See Section 4 for further discussion of CS dimensionality.)

Before proceeding, several assumptions commonly used in FST are listed. First, the functions

$$P_i = R_{u_i} \qquad [16]$$

and

$$\begin{aligned} P_i(t) &= \text{probability that item } \underline{i} \text{ is answered correctly given an} \\ &\quad \text{examinee with ability equal to } \underline{t} \text{ has been sampled} \\ &= E(u_i|\theta = t) \qquad [17] \end{aligned}$$

are assumed to be continuous. In addition, all abilities are assumed to lie in a closed bounded interval I. The assumption of continuous P_i is restrictive and may have to be dropped for some applications. The assumption of bounded abilities, on the other hand, results in no loss of generality because any latent trait model can be reformulated by routine methods as an isomorphic model with bounded abilities. These assumptions together imply that the canonical space consists of continuous functions on the interval I.

A major result of FST is that two densities are equivalent in the sense of problem 1 above if they have the same projection into the canonical space. Thus, if some J functions $h_1, h_2, \ldots, h_J$ form a basis for the CS and if

$$\int_I h_j(t) f_1(t)dt = \int_I h_j(t) f_2(t)dt \qquad j = 1,2, \ldots, J \qquad [18]$$

then there is no objective way to choose between f_1 and f_2.

By an "objective way to choose between f_1 and f_2" is meant a method of using the observables (the item scores) to decide which of f_1 or f_2 is more nearly correct. This is impossible because it can be proven that every statistic has the same probability distribution when f_1 is correct as when f_2 is correct.

This fact leads to a useful representation of densities. An arbitrary density *f* can be represented uniquely as

$$f(\cdot) = f_0(\cdot) + \sum_{j=1}^{J} \alpha_j h_j(\cdot)$$

$$= f_0(\cdot) + f^*(\cdot) \qquad [19]$$

where $\{h_j\}$ is a basis for the CS and f_o is orthogonal to the regression function of every statistic, i.e.,

$$\int_I E(S|\theta = t) f_0(t)dt = 0 \qquad [20]$$

for every statistic S. In this decomposition, f_o is called the *null* part of *f*, f* the *identifiable* part of *f*, and α_j the *j*th *coordinate* of *f*.

The identifiable part of *f* is indeed identifiable because a sequence of estimators $\{\hat{f}_N(t)\}$ can be constructed that will (almost surely) converge to $f^*(t)$ as sample size N is increased. The convergence turns out to be uniform in *t*.

The null part of *f* is null in the sense that it is totally unrelated to data. There is no objective way to use the administered items to distinguish two densities with the same identifiable parts and different null parts. Such densities cannot and, for most puposes, need not be distinguished. Both densities lead to the same predictions in all applications. A proposition or scientific statement that is true if f_1 is the ability density and false if f_2 is the ability density is (in the technical foundations-of-measurement sense of the term) not meaningful. The preposition may be interesting, clearly stated and important, but there will be no way to tell if it is true or false from the observed responses to the test items.

The representation leads to a strategy for estimating densities. If the h_j (called *coordinate functions*) are orthonormal, then the coordinates α_j have a statistical interpretation,

$$\alpha_j = E[h_j(\theta)] \ , \qquad [21]$$

i.e., α_j is the expected value of a function of the unobserved ability θ.

Since h_j is in CS, h_j is the regression function of some statistic, say X_j, and its regression function, the conditional expected value of X_j, will be equal to h_j:

$$E(X_j|\theta = t) = h_j(t) \ . \qquad [22]$$

Since X_j is simply a function of item scores, X_j can be computed for each examinee in a large sample of, say, N examinees to obtain a sample mean $\overline{X}_j$. By the law of large numbers the estimate $\hat{f}_N(t)$

$$\hat{f}_N(t) = \sum_{j=1}^{J} \overline{X}_j h_j(t) \qquad [23]$$

will converge (in probability) to the identifiable part of f as the sample size N becomes large.

This estimate is especially well behaved and easy to study because the examinees are independently sampled. In fact, for sufficiently large sample size N the vector of sample averages $N^{1/2}\langle\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_J\rangle$ will be nearly multivariate normal with covariance matrix that, at least in some applications, can be regarded as known.

3. An Outline of Some Basic Theory

For clarity and ease of future reference, Section 2 is outlined with the introduction of some additional assumptions and notation.

Basic Latent Trait Model and Notation (Simplest One-Dimensional Version)

$\Omega = \{\omega\}$ = the probability space
= an infinite set of actual or conceivable examinees available for sampling and testing
θ = ability random variable. $\theta(\omega)$ is unobserved.
$f(\cdot)$ = the density for θ. Its support will always be assumed to be contained in an interval I. Except when noted, only continuous densities are considered.
I = a closed interval containing all abilities:

$$\int_I f(t)dt = 1 . \qquad [24]$$

n = number of test items.
U = a random n-vector of item scores (the only observables)
= $\langle u_1, u_2, \ldots, u_n\rangle$.
u_i = item score random variable. $u_i(\omega)$ is either zero or one.
$P_i(\cdot)$ = item response function, or item characteristic function.

$$P_i(t) = E(u_i|\theta = t) . \qquad [25]$$

These functions are assumed to be continuous.
Local independence assumption: For any n vector of zeros and ones

$$U^* = \langle u_1^*, u_2^*, \ldots, u_n^*\rangle , \qquad [26]$$

$$\mathrm{Prob}\{U = U^* | \theta = t\} = \prod_{i=1}^{n} \{u_i^* P_i(t) + (1 - u_i^*)[1 - P_i(t)]\} \qquad [27]$$

Statistic: A number-valued function of the item scores.

Basic Formula Score Terminology

Regression Function: The regression function of a statistic S is the conditional expectation

$$R_S(t) = E[S|\theta = t] \; . \qquad [28]$$

Canonical Space (CS): The real vector space of all regression functions for a test. It is a finite dimensional subspace of the vector space of all continuous functions defined on I and can be shown to have dimension $\leq 2^n$.

J: The dimension of the canonical space (discussed in Section 4).

$(\cdot,\cdot)$: Notation for the inner product used on the space of continuous functions defined on I.

$$(g,h) = \int_I g(t)h(t)dt \; . \qquad [29]$$

Note that

$$(h,f) = E[h(\theta)] \; . \qquad [30]$$

Coordinate Functions: An orthonormal basis for the CS of a test, generally denoted $\{h_1, h_2, \ldots, h_J\}$.

α_j: The projection of the ability density on the jth coordinate function, h_j. It is called the jth coordinate of f and has statistical interpretation

$$\alpha_j = E[h_j(\theta)] \; . \qquad [31]$$

Ability Density Equivalence

$\chi_A(\cdot)$: The indicator function of the Set A

$$\chi_A(t) = 0 \text{ if } t \text{ is not in } A$$
$$= 1 \text{ if } t \text{ is in } A \; . \qquad [32]$$

P[.;S,.]: Notation for the probability distribution of the statistic S. $P[A;S,f_1]$ is the probability that the statistic S is in the Set A when f_1 is the density for θ.

$$P[A;S,f_1] = \int E[\chi_A(S)|\theta = t]f_1(t)dt \qquad [33]$$

Equivalent Densities: Two densities f_1, f_2 are equivalent for every statistic S

$$P[\cdot;S,f_1] = P[\cdot;S,f_2] \qquad [34]$$

i.e., if the probability distribution of each statistic is the same when $f = f_1$ as when $f = f_2$.

Characterization of Equivalent Ability Densities: f_1 is equivalent to f_2 if and only if

$$(f_1,h_j) = (f_2,h_j) \text{ for } j = 1,2, \ldots, J \qquad [35]$$

for any set of coordinate functions $\{h_j\}$.

Ability Density Decomposition

$g = g_o + g^*$: Every density g on I can be expressed uniquely in the form:

$$g(\cdot) = g_0(\cdot) + g^*(\cdot) \qquad [36]$$

where g^* is in the canonical space and for every statistic S

$$(R_S,g) = (R_S,g^*) \qquad [37]$$

g^*: The identifiable part of a density g. The projection of the density into the CS.

$$g^*(t) = \sum_{j=1}^{J} \alpha_j h_j(t) \qquad [38]$$

for coordinate functions $h_1, h_2, \ldots, h_J$.

g_o: The null part of a density g

$$g_0 = g - g^* . \qquad [39]$$

cannot and generally need not be estimated because $f_1^* = f_2^*$ implies $P[\cdot;S,f_1] = P[\cdot;S,f_2]$ for every statistic S.

Ability Density Representation

Densities and J vectors: The mapping

$$g \rightarrow \langle (g,h_1),(g,h_2), \ldots, (g,h_J) \rangle \qquad [40]$$

associates each density on I with a unique J vector.
Densities associated with the same J vector are equivalent.

Consistent, Unbiased Estimates of the Identifiable Part of the Ability Density

X_j: A statistic with regression function equal to h_j, that is,

$$R_{X_j}(\cdot) = h_j(\cdot)\,. \tag{41}$$

There must be at least one because h_j is in the CS, and the CS consists of regression functions only. X_j must be bounded because it has all of its probability on a set of 2^n points.

$E(X_j) = \alpha_j$: Follows from

$$\begin{aligned} E(X_j) &= E[E(X_j|\theta)] \\ &= E[R_{X_j}(\theta)] \\ &= E[h_j(\theta)] \\ &= (h_j, f) \end{aligned} \tag{42}$$

$\bar{X}_{j;N}$: Sample mean of X_j from a sample of N examinees.

$\bar{X}_N = \langle \bar{X}_{1,N}, \bar{X}_{2,N}, \ldots, \bar{X}_{J,N} \rangle$: Sample mean of N bounded independent, identically distributed random vectors. Converges to $\langle \alpha_1, \alpha_2, \ldots, \alpha_J \rangle$. Asymptotically, $N^{1/2}\bar{X}_N$ is multivariate normal.

$\hat{f}_N(t)$: $$\hat{f}_N(t) = \sum_{j=1}^{N} \bar{X}_{j,N} h_j(t) \tag{43}$$

a consistent, unbiased estimate of the identifiable part of the ability density. Asymptotically, $N^{1/2}\hat{f}_N(t)$ is normal.

Construction of Coordinate Functions h_j and Coordinate Estimators X_j: See Section 4.

4. Implementation of Ability Density Results

In this section the abstract results given in Sections 2 and 3 will be applied to estimate densities and ICCs. This section is included to show in a general way how the theory is used to analyze data. The discussion is organized about four technical questions that commonly arise in response to presentations of the theory:

1. How are the coordinate functions $h_1, h_2, \ldots, h_J$ determined in actual applications?
2. How are the statistics X_j specified?
3. What is the dimension J and how is it determined?

BEST COPY AVAILABLE

4. Can the calculations be arranged in a way to avoid very long, numerically unstable calculations?

A set of statistics $\{S_k\}$ is called complete if any statistic S can be written as a linear combination of finitely many of them. For example, the elementary formula scores $\{v_k\}$ formed by considering all products of item scores are complete. These are the scores:

$$\begin{array}{l} 1 \\ u_1, u_2, \ldots, u_n \\ u_1u_2, \ldots, u_{n-1}u_n \\ \vdots \\ u_1u_2 \ldots u_n . \end{array} \tag{44}$$

The regression functions

$$R_{v_k}(t) = E[v_k | \theta = t] \tag{45}$$

are simply products of the ICCs P_i.

A set of coordinate functions can be constructed from any complete set of statistics as follows. First a function on I × I is specified by the formula

$$H(s,t) = \sum_k R_{S_k}(s)\, R_{S_k}(t) . \tag{46}$$

"Is specified" in practical terms means that a computer subroutine is written that accepts pairs of numbers s,t as input and returns the number given on the right-hand side as output. In the special case where $\{S_k\} = \{v_k\}$, the "elementary" formula scores, it can be shown that H has the easily calculated form

$$H(s,t) = \prod_{i=1}^{n} [1 + P_i(s)P_i(t)] . \tag{47}$$

This special case will be discussed after the general case is treated.

Using standard methods H(s,t) is decomposed into a finite sum of products of orthonormal functions. More specifically, a set of positive numbers $\lambda_1 \geq \lambda_2 \geq, \ldots, \lambda_J > 0$ and orthogonal functions h_j, $j = 1, 2, \ldots, J$

$$(h_j, h_{j'}) = \begin{cases} 1 \text{ if } j = j' \\ 0 \text{ otherwise} \end{cases} \tag{48}$$

are computed such that

$$H(s,t) = \sum_{j=1}^{J} \lambda_j h_j(s) h_j(t) \qquad [49]$$

for all s, t in I. Just as the eigenvalues of a positive semidefinite matrix are determined by the matrix, the λ's and the "rank" J are determined by H. It can be shown that every function in the CS is a linear combination of the h_j no matter which complete set of statistics is used to construct H. In other words, the h_j are coordinate functions, and J is the dimension of the CS.

To construct an estimator of the coordinate

$$\alpha_j = E[h_j(\theta)] \ , \qquad [50]$$

note that h_j is an eigenfunction of the "linear operator" defined by H. In symbols,

$$\begin{aligned} \lambda_j h_j(t) &= \int H(s,t) h_j(s) \, ds \\ &= \sum_k R_{S_k}(t) \int R_{S_k}(s) \, h_j(s) \, ds \\ &= \sum_k R_{S_k}(t) \, (R_{S_k}, h_j) \ . \end{aligned} \qquad [51]$$

If

$$R_{S_k}(t) = E[S_k | \theta = t] \qquad [52]$$

is replaced by S_k/λ_j in this formula, a statistic X_j is specified:

$$X_j = \lambda_j^{-1} \sum_k S_k (R_{S_k}, h_j) \ . \qquad [53]$$

Since $E(S_k | \theta = t)$ is $R_{S_k}(t)$, it follows that

$$E(X_j | \theta = t) = h_j(t) \qquad [54]$$

and

$$E(X_j) = E[h_j(\theta)] \ . \qquad [55]$$

Thus, the sample mean $\overline{X}_j$ is a consistent, unbiased estimator of the coordinate α_j.

J, the dimensionality of the CS of the test, was calculated by analyzing any complete set of statistics $\{S_k\}$. J turned out to be the "rank" of the linear operator

$$h \to \varphi(h) = \int \sum_k R_{S_k}(\cdot) R_{S_k}(t) h(t) dt = \int H(\cdot,t) h(t) dt \ . \qquad [56]$$

Although the eigenfunctions h_j and the eigenvalues λ_j depend on the choice of $\{S_k\}$, J does not. However, some applications may lead to particular $\{S_k\}$ and subsequently to a decision to treat the CS as if it had dimension $J' < J$.

The typical situation arises when the problem is considered of selecting $\hat{f}$ so as to minimize a quadratic index of goodness of fit, such as

$$Q(\hat{f}) = \sum_k [\bar{S}_k - E(S_k;\hat{f})]^2 \ . \qquad [57]$$

Here $\bar{S}_k$ is the sample average value of S_k and

$$E(S_k;\hat{f}) = \int_I R_{S_k}(t) \hat{f}(t) dt \qquad [58]$$

is the predicted value of $\bar{S}_k$. If

$$H(s,t) = \sum_{j=1}^{J} \lambda_j h_j(s) h_j(t) \quad , \qquad [59]$$

then Q can be written in the form

$$Q(\hat{f}) = \sum_{j=1}^{J} \lambda_j [\hat{\alpha}_j - \bar{X}_j]^2 + \text{terms that are independent of } \hat{f} \qquad [60]$$

where

$$\hat{\alpha}_j = (\hat{f}, h_j) \qquad [61]$$

$$\bar{X}_j = \lambda_j^{-1} \sum_k \bar{S}_k (R_{S_k}, h_j) \ . \qquad [62]$$

From Equation 60 it can be seen that the size of λ_j measures the degree of improvement of fit to a set of statistics $\{S_k\}$ that can be obtained by including one more term in the representation of the identifiable part of the ability density

$$\sum_{j'<j} \alpha_{j'} h_{j'}(\cdot) \ . \qquad [63]$$

If λ_j is very small or if $\bar{X}_j$ has very large sampling error, then we do not attempt to estimate the coordinate and proceed as if the canonical space has lower dimensionality than J.

In applications J' has been selected by computing the λ_j and treating very small λ_j's as zero. As a check on the adequacy of this procedure the differences between

$$\sum_{j=1}^{J'} (g,h_j)h_j(\cdot) \qquad [64]$$

and $g(\cdot)$ are examined for various functions g. The functions g generally considered are the P_i, selected regression functions, and various guesses about $f(\cdot)$. The two functions will agree exactly if g is in the CS and J' is the dimensionality of the CS.

One formula scoring technique has proven more accurate than all of the other techniques that have been tried. The complete formula score method (CFSM) uses the elementary formula scores $\{v_k\}$ as its complete set of statistics and begins with the identity

$$H(s,t) = \sum_k R_{v_k}(s)R_{v_k}(t)$$

$$= \prod_{i=1}^{n} [1 + P_i(s)P_i(t)] \; . \qquad [65]$$

(The identity is verified by induction or by expanding the product.) The sum has 2^n terms, but the product has only n terms and thus can be calculated with many fewer operations.

The X_j can also be calculated with a variant of this identity. Each sampled examinee's data is transformed to define a random continuous function $V(t)$, which is called the V-transform of the examinee's data. This function

$$V(t) = \prod_{i=1}^{n} [1 + u_i P_i(t)] \qquad [66]$$

is easily calculated and is identically equal to

$$\sum_{k=1}^{2^n} v_k R_{v_k}(t) \; . \qquad [67]$$

Therefore, $\int_I V(t)h_j(t)dt$ equals

$$\sum_k v_k(R_{v_k},h_j) = \lambda_j X_j \; . \qquad [68]$$

This reduces the calculation of X_j from 2^n operations to one numerical integration. In fact, to calculate the sample mean $\overline{X}_j$, only one integration need be done. This is true because the sample average of the integrated V(t) is the integral of the sample average

$$\text{Ave}\{\int V(t)h_j(t)dt\} = \int[\text{Ave } V(t)]h_j(t)dt \ . \qquad [69]$$

Thus, in applications, V(t) is computed on a grid of t values for each examinee, accumulated over examinees and numerically integrated to obtain the J sample means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_J$. This procedure can be adapted to compute sample covariances of the X_j by numerical integration.

5. Discovering the Shapes of Item Response Functions

Sometimes, but not always, the shape of ICCs can be rationally deduced and parameterized. For example, the S-shape of the 3-parameter logistic curve on an unbounded ability continuum follows from

1. Montonicity. More able examinees are more likely to answer correctly.
2. Asymptotes. Although the probability of a correct response can be made arbitrarily close to 1.0 by sampling from very high ability subpopulations, a substantial proportion of each low ability subpopulation will select the correct option of a well-constructed multiple-choice item.
3. Simplicity/Parsimony. The item response curve has no more points of inflection than the one implied by the monotonicity and asymptotes conditions above and by smoothness conditions.
4. Symmetry. The graph of the ICC is symmetric for high and low ability levels in the sense that a length-preserving transformation $(x,y) \rightarrow (2x_0-x, 2y_0-y)$ about a point of inflection on the graph (x_0,y_0) carries the curve into itself.

Fitting 3-parameter logistic functions is sensible when these conditions are met because every curve satisfying these conditions will be close to at least one 3-parameter logistic function.

Sometimes these assumptions are implausible or clearly false, and a method is needed to discover and parameterize shape. As a one-dimensional example, it was demonstrated (Levine & Drasgow, in press) with a very large sample of aptitude test examinees that the conditional response function for the response of choosing a particular (incorrect) option on a multiple-choice test for several items was clearly nonmonotonic. In multidimensional measurement the shape of the item characteristic surface is a matter of considerable psychological importance because it represents a statement about how several abilities interact to simultaneously determine response probability. In the next section, the methods discussed here are used to develop a procedure for determining the shape of an item characteristic surface.

Another application of the method described in this section is the study of item responses, such as omitting, that cannot reasonably be expected to satisfy local independence. FST permits the construction of a consistent estimate of an "omitting characteristic curve" $P\{\text{item } i \text{ is omitted} \mid \theta = t\}$, without assuming local independence for omitting responses.

The basic issues addressed in this segment of the research are the following:

1. What is the shape of a new item or item type´s ICC?
2. What information about ability is contained in incorrect answers and item omitting?
3. Modeling item responses that may not be locally independent, such as omitting and the following examples drawn from computer-administered tests: (a) attempting to change an answer, (b) requesting a display of a previously presented item or part of an item, (c) responding in a time clearly too short to read the item, (d) responding with potentially damaging force to the terminal.

Equations Relating FST to ICC Shape

Our results in this area depend on the following elementary relation which is used to reduce IRF estimation to ability density estimation:

$$\text{Prob}\{\text{correct response} \mid \text{ability is in Set A}\}$$

$$= \frac{\text{Prob}\{\text{ability is in A} \mid \text{correct response}\}}{\text{Prob}\{\text{ability is in A}\}} \times \text{Prob}\{\text{correct response}\} \qquad [70]$$

Thus, the conditional response probability is proportional to the ratio of the ability distribution in the subpopulation of examinees correctly answering the item to the unconditional distribution. (The constant of proportionality, Prob{correct response} = $\bar{P}$, has an obvious consistent unbiased estimate, the sample proportion correct). If regularity assumptions are made, then

$$P(t) = \frac{f^{+}(t)}{f(t)}\,\bar{P} \qquad [71]$$

where f^{+} is the ability density in the subpopulation of item passers and f is the ability density. This equation can be written in the form

$$P(t) = \frac{1}{1 + \dfrac{\bar{Q}}{\bar{P}}\,\dfrac{f^{-}(t)}{f^{+}(t)}} \qquad [72]$$

where $\bar{Q} = 1 - \bar{P}$ and f^{-} is the failure density.

This formula is especially useful for developing a distribution theory for the estimates because the failure and passing subpopulations are disjoint. Thus, the distribution of an estimator of $f^{+}(t)/f^{-}(t)$ can be developed by studying the ratio of statistically independent random variables.

Several variations of this approach to ICC estimation have been tried with generally satisfactory but occasionally poor results. Systematic comparisons of the variations will be made after more exploratory work. Some current and projected refinements are delineated below.

1. If ability is uniformly distributed or if $f(\cdot)$ is constant on a range of abilities of interest, then the ICC is proportional to f^+, and ICC estimation is simplified. Furthermore, sampling fluctuation is unlikely to give a zero or negative density estimate. An initial approximation can be used to transform ability so that $f(\cdot)$ is approximately constant.
2. The sampling distribution of coordinate estimates depends on the choice of $\{S_k\}$. It has been observed that if h_1, the coordinate function with the largest eigenvalue is close to $f(\cdot)$, then very good estimates of $f(\cdot)$ are obtained. Currently, an attempt is being made to capitalize on this effect by carefully choosing $\{S_k\}$ and controlling the eigenfunction shapes.
3. The current density estimates are least squares in the sense that they minimize the residual sum of squares Q (Section 4). The results on density equivalence permit the expression of the likelihood function as a function of finitely many parameters, the coordinates α_j. In principle, this expression could be maximized and maximum likelihood density estimates could be computed.

The basic equation used to relate ICCs to density ratios is essentially the definition of conditional probability. Local independence plays no role in its derivation. In fact, the binary item response on the focal item is being used merely to divide the sample into those who answer the item correctly and those who answer it incorrectly. The equation would remain valid if any binary score were used to dichotomize the sample and population. Therefore, the same estimation techniques used to estimate ICCs could be used to study the relation of ability to complex item responses that failed to satisfy the local independence assumption. These include item skipping and very fast responding.

6. Multidimensional Formula Scoring for Homogeneous Subtests

There is an interesting and important multidimensional measurement problem that can be implemented with currently available unidimensional software. The problem is to discover how several abilities jointly determine response probability on new item types. This can be done when certain psychological assumptions are valid. In particular, it is necessary that a variety of item types are available, that all items depend on the same small number of abilities, and that items can be grouped into "homogeneous" subtests.

For concreteness consider three item types: synonyms, antonyms, and analogies. Suppose that all three depend only on a pair of abilities θ_1, θ_2 in the sense that for any ability levels s, t and any r items, the item scores u_{i_1}, u_{i_1}, ..., u_{i_r} satisfy

$$E\left[\prod_{j=1}^{r} u_{i_j} \,\middle|\, \theta_1 = s \;\&\; \theta_2 = t\right] = \prod_{j=1}^{r} E\left[u_{i_j} \,\middle|\, \theta_1 = s \;\&\; \theta_2 = t\right] . \qquad [73]$$

In other words, the items are independent in subpopulations formed by conditioning on both abilities.

Homogeneous subtests, defined more formally below, are unidimensional subtests of a multidimensional test. For example, both synonym and antonym items may require both language fluency, θ_1, and an ability to recognize and generalize abstract relations, θ_2; but the antonym items can be written in such a way as to demand a relatively large amount of the second ability. Thus, synonym items may satisfy local independence with respect to some linear or nonlinear function of θ_1 and θ_2, say $\theta_1 + \theta_2$, and antonym items with respect to, say $\theta_1 + 2\theta_2$. Subtests consisting of one item only will appear unidimensional, but the total test will not.

Note that the assumption that item types from homogeneous subtests is more general (and more believable) than the assumption that different item types measure different traits. This should be obvious after the discussion of "ad hoc coordinates."

If certain plausible assumptions (described below) are correct, then unidimensional parameter estimation programs can be used with a test consisting only of synonym items and a test consisting only of antonym items. FST can then be used to represent and to estimate bivariate analogy IRFs.

Homogeneous Subtests and Ad Hoc Coordinates

Latent trait theory provides a way to precisely state what is meant by a homogeneous subtest and items requiring different amounts of unobserved abilities. As before, the population of examinees is denoted by a point set $\Omega = \{\omega\}$. In this situation, abilities map examinees into two vectors rather than numbers: $\theta(\omega) = \langle\theta_1(\omega), \theta_2(\omega)\rangle$. Because examinees are randomly sampled, θ is a random vector.

To quantify the notion of homogeneous subtests, a pair of number-valued functions ϕ, ψ are considered. The first subtest is homogeneous in the sense that it satisfies a local independence assumption relative to $\phi(\theta)$. In symbols, for items i_1, i_2, ..., i_r on the first subtest

$$\text{Prob}\{u_{i_1} = 1, u_{i_2} = 1, \ldots \text{ and } u_{i_r} = 1 \mid \phi(\theta) = t\} = \prod_{j=1}^{r} P_{i_j}(t) \qquad [74]$$

where

$$P_{i_j}(t) = \text{Prob}\{u_{i_j} = 1 \mid \phi(\theta) = t\} . \qquad [75]$$

A similar equation expresses the assumption that the second subtest is homogeneous with respect to $\psi(\theta)$: The item scores for items in the second subtest are independent in the subpopulation of examinees with the property $\psi(\theta) = s$ for each constant s.

Note that these conditions permit using available parameter estimation programs to calibrate each subtest separately. Somewhat paradoxically, each item is essentially multidimensional, but each subtest satisfies the axioms of one-dimensional latent trait theory. FST provides a method for integrating the subtests and modeling the analogy items that also depend on fluency and abstraction, but to an unknown extent.

The functions ϕ and ψ are called "ad hoc coordinates." If for each $\langle s,t\rangle$ there is at most one vector $\langle x,y\rangle$ satisfying $s = \phi(x,y)$ and $t = \psi(x,y)$, and certain regularity conditions are met, then ϕ and ψ can be treated as curvilinear coordinates for the set of (bivariate) abilities.

The FST analysis to be presented only gives a representation of item response function in terms of ad hoc coordinates

$$P_i(s,t) = \mathrm{Prob}\{u_i = 1 | \phi(\theta) = s \ \& \ \psi(\theta) = t\} \qquad [76]$$

For many modeling problems this is adequate. Admittedly, the variables θ_1 (fluency) and θ_2 (abstraction) are considerably more interesting than ϕ and ψ. Conjoint measurement or uniform systems analysis (Levine, 1970) may permit the analysis of the relation between ϕ,ψ and θ_1,θ_2. However, the current concern is with the psychometric problem of representing new items in the ad hoc system.

Formula Score Approach to Representing New Item Types

Consider a 21-item test consisting of 10 synonym items followed by 10 antonym items and one analogy item. The task is to compute

$$\mathrm{Prob}\{u_{21} = 1 | \phi(\theta) = s \ \& \ \psi(\theta) = t\} = P(s,t) \ . \qquad [77]$$

First, the homogeneous subtests will be analyzed separately. The most direct approach would be to first calibrate the synonym items by embedding them in a large conventional administration of many items of the same type. The analysis would yield ϕ ICCs

$$\mathrm{Prob}\{u_i = 1 | \phi(\theta_1,\theta_2) = t\} = P_i(t) \qquad i = 1,2, \ldots, 10 \ . \qquad [78]$$

Similarly, a separate anlysis of the antonym items yields ψ ICCs

$$\mathrm{Prob}\{u_i = 1 | \psi(\theta_1,\theta_2) = t\} = P_i(t) \qquad i = 11,12, \ldots, 20 \ . \qquad [79]$$

To discover the shape of an analogy item's ICC, a 21-item test would be administered to a sample of examinees. By an obvious generalization in Section 5 the analogy item ICC

$$P(s,t) = \mathrm{Prob}\{u_{21} = 1 | \phi(\theta_1,\theta_2) = s \text{ and } \psi(\theta_1,\theta_2) = t\} \qquad [80]$$

can be represented as a ratio of densities.

$$P(s,t) = \frac{f^{+}(s,t)}{f\ (s,t)}\,\bar{P} \qquad [81]$$

where f^+ is the (bivariate) density in the subpopulation of examinees who correctly answered Item 21 and f is the unconditional density.

Before continuing this analysis, the CS for the 20-item test is described. It turns out that the assumptions imply that the canonical space of the 20-item test is simply the "tensor product" of the CS for the first subtest and the second subtest.

At this point it seems advisable to restate some definitions. The CS for the first subtest is the set of one-dimensional regression functions

$$E[S|\phi(\theta_1,\theta_2) = t] = R_S(t) \qquad [82]$$

where S is a statistic whose value depends on the first 10 item scores only. The CS for the antonym item is the set of regression functions $E[S|\psi(\theta_1,\theta_2) = t]$ for statistics S that are functions of the second 10 scores only.

The CS for the first 20-item subtest will be the set of regression functions

$$R_S(s,t) = E[S|\phi(\theta_1,\theta_2)] = s \qquad [83]$$

$$\psi(\theta_1,\theta_2) = t \qquad [84]$$

for the statistics S of the first 20 scores. The psychometric assumptions imply that any $R(s,t)$ in the 20-item CS can be written as a finite sum of functions of the form $h(s)h'(t)$ for h in the first subtest CS and h' in the second subtest CS. In fact if $\{h_1, h_2, \ldots, h_J\}$ is a basis for the first CS and $\{h'_1, h'_2, \ldots, h'_{J'}\}$ for the second CS, then the $J \times J'$ functions

$$h_j(s)h'_{j'}(t) = h_{jj'}(s,t) \qquad [85]$$

where $j = 1, 2, \ldots, J$ and $j' = 1, 2, \ldots, J'$ will be a basis for the two-dimensional CS.

Current one-dimensional FS programs can be used to estimate the bivariate densities f^+ and f. The estimate will have form

$$f^+(s,t) = \sum_{j=1}^{J} \sum_{j'=1}^{J'} \bar{X}_{jj'} h_{jj'}(s,t) \qquad [86]$$

where the sample mean $\bar{X}_{jj'}$ is a consistent, unbiased estimator of $E[h_{jj'}(\theta)]$. If f^+ and f are in the 20-item CS, then (by sample splitting to estimate f^+ and f separately) a consistent estimate of $P(s,t)$ is easily specified, and the shape of the item response surface can be "discovered."

The success of this approach requires f^+ and f to be in or near the 20-item two-dimensional CS. This assumption seems plausible when the variety of shapes that can be constructed as linear combinations of the $J \times J'$ coordinate functions are considered.

6. Are Two Tests Measuring the Same Trait(s)?

Suppose a major revision is made of a complex, not necessarily unidimensional test. Does the new test measure the same traits? Suppose the format or mode of administration is changed. Does the test still measure the same traits? Suppose a translation of the test into another language is attempted and that it is unlikely that every original test item is psychometrically equivalent to its translation. Can the translated test nonetheless measure the same traits as the original? These questions lead to asking the foundations-of-measurement question,

> What necessary and/or sufficient conditions must item scores
> obey before it can be concluded that two nonparallel tests
> are measuring the same trait(s)?

It is hoped that theoretical work on this problem will lead to a statistical test that can be used in applications. This section relates FST to the problem and reports some current work.

Two nonparallel tests are administered to the same population. (The tests can be considered as subtests of one test.) Item response curves are fitted to each test separately. In other words, except possibly to compute an equating transformation, only Test 1 scores are used to estimate the IRF of a Test 1 item. Can the variable in the first set of IRFs be given the same interpretation as the variable in the second set of IRFs?

The mathematical kernel of this problem seems to be this. A probability measure is given for a set Ω along with two sets of zero-one random variables: $u_1, u_2, \ldots, u_n$ and $u_1', u_2', \ldots, u_{n'}'$. Two sets of real functions, $P_1, P_2, \ldots, P_n$ and $P_1', P_2', \ldots, P_{n'}'$, are also given. What conditions must be assumed in order for it to be possible to construct one more random variable θ such that the P's are IRFs for the $\underline{u}'$s relative to θ and all $n + n'$ $\underline{u}'$s are locally independent relative to θ.

A strong necessary condition on the given scores and functions can be formulated with FS notation and concepts. Let $\{v_k\}$ denote the elementary scores in the first set of $\underline{u}'$s (Section 5). Let $\{v_{k'}'\}$ denote the elementary scores in the second set of $\underline{u}'$s. Let R_{v_k} and $R_{v_{k'}'}$ be the corresponding products of P's. Thus, if

$$v_k = u_{i_1} u_{i_2} \cdot \cdot \cdot u_{i_r} \qquad [87]$$

then

$$R_{v_k}(\cdot) = P_{i_1}(\cdot) P_{i_2}(\cdot) \cdot \cdot \cdot P_{i_r}(\cdot) \ . \qquad [88]$$

Let $\{v_{k,k'}\}$ denote the elementary formula scores for the $n + n'$ item test where $v_{k,k'}$ is $v_k v'_{k'}$. Let $R_{k,k'}$ denote the corresponding product of functions for $v_{k,k'}$, that is,

$$R_{k,k'} = R_{v_k} R_{v_{k'}'} \ . \qquad [89]$$

Let $\{h_j\}_{j=1}^{J}$ be an orthonormal basis for the linear span of the $\{R_{k,k'}\}$. For example, the h_j could be obtained by analyzing the function $H(s,t)$ defined by

$$\prod_{i=1}^{n} [1 + P_i(s)P_i(t)] \prod_{i=1}^{n'} [1 + P_i'(s)P_i'(t)] \qquad [90]$$

as in Section 4. Finally, let X_j be the random variable

$$X_j = \sum_k \sum_{k'} v_{k,k'} (R_{k,k'}, h_j) \ , \qquad [91]$$

which can be more conveniently computed as described in Section 4.

It is easy to show the following condition on expected values of scores

$$E(v_k v_{k'}) = \sum_{j=1}^{J} E(X_j)(h_j, R_{k,k'}) \qquad [92]$$

is necessary for all elementary formula scores $v_k v_{k'}$.

The condition also appears to be sufficient, although a proof is not available at this time. In any event, additional work is needed to determine when one random variable suffices for Ω and specified subpopulations of Ω.

ACKNOWLEDGMENTS

This research was supported by Office of Naval Research Contract N00014-79C-0752, NR 150-445.

REFERENCES

Johnson, N.L., & Kotz, S. Continuous univariate distributions (Vol.2). Boston: Houghton-Mifflin, 1970.

Levine, M. V. Transformations that render curves parallel. Journal of Mathematical Psychology, 1970, 7, 410-443.

Levine, M. V. Tests of local independence and item fit (Working Paper 81-1). Champaign-Urbana: University of Illinois, Department of Educational Psychology, 1981.

Levine, M. V., & Drasgow, F. The relation between option choice and estimated ability. Educational and Psychological Measurement, in press.

DISCUSSION

ROBERT J. MISLEVY
INTERNATIONAL EDUCATIONAL SERVICES

Before the days of item response theory, test data were analyzed in terms of indices such as percent-correct and item-test correlations, which jointly describe the interactions of a sample of persons and a collection of test items. IRT provides an escape from these sample-bound statistics by breaking the problem into pieces, with the assumption that item responses can be explained in terms of conceptually distinct characteristics of persons and items. The key concept is the item response curve, or item response function, which gives the probability of a correct response as a function of person ability--independent of how many people happen to be at that level of ability.

Much research has focused on the pieces in this problem that at least conceptually, and occasionally algebraically (Rasch, 1960) do not depend on the distribution of person ability parameters, particularly the estimation of item response functions and the estimation of a single person's ability from his or her pattern of responses. Less attention has been accorded the piece that remains, namely, the distribution of ability parameters in a sample of persons. As Levine has pointed out, however, there are a number of problems in applied work with item response models for which this is exactly the sort of information needed.

Of course, the distribution of ability could always be approximated by the distribution of ability estimates. This approach is not very satisfactory, however, since the two distributions can differ substantially, especially for short tests. The distribution of ability estimates, for example, will always exhibit greater variation than the underlying true distribution due to the presence of measurement error variation. More satisfactory procedures for estimating a latent ability distribution (e.g., Andersen & Madsen, 1977; Lord, 1969; Sanathanan & Blumenthal, 1978) have been based on the expression for the marginal probability of each score pattern $\underline{x}$ as a function of the parameters ξ of an underlying ability distribution $\underline{g}$ of known parametric form:

$$P(x|\xi) = \int_{\theta} P(x|\theta)\ g(\theta|\xi)\ d\theta. \qquad [1]$$

Lord has estimated $\underline{g}$ by equating moments of the observed score distribution as predicted by this convolution formula to those of the sample; the other authors find values of ξ that maximize the likelihood of the observed data.

Levine's approach to the problem of estimating a latent distribution differs radically. What he has done, in essence, is to go back to the other pieces of the problem, namely, the item response functions, and to use them as building blocks for a new vector space--his "canonical space" of a test, which consists of all the item response functions, the functions given by their pointwise averages and weighted averages, their products, products of averages and averages of products, and so on. He has demonstrated (1) that this space can usually be represented by a small number of basis functions and (2) that the member of this space (as given by the weighted average of a set of basis vectors) that best approximates the latent distribution of ability in a sample of persons can be found in a relatively straightforward manner.

Compared with the marginal solution, the formula score solution has both advantages and disadvantages. One great advantage is its (relative) simplicity; a great deal of heavy iterative computation and numerical integration can be avoided. A second advantage is in having what could be called "virtually sufficient" statistics for ability parameters. Whereas the form of Rasch models guarantee a priori that total scores contain all information about abilities, the formula score procedure can be applied to item response models of any form to determine which n weighted combinations of item scores convey nearly all the information about abilities that can be determined from responses to the items that comprise the test.

One disadvantage of the procedure is that although there is a guarantee of finding the member of the canonical space that best approximates the latent distribution of ability, how good this approximation will be is unknown. It would seem to depend on the number and variety of response functions there are to work with; some theorems about the density of canonical spaces would be desirable. Just as it is known that any continuous function can be approximated by a polynomial of a sufficiently high degree, it could turn out, for example, that any ability distribution can be approximated by a test with sufficiently many equally spaced items with parallel response functions. Similarly, it would be desirable to know the "robustness" of the virtually sufficient statistics; does substituting one or two items on a test yield substantially different optimal formula scores, in form and/or in number?

I have mentioned a number of advantages of the formula score approach to estimating latent distributions, along with some suggestions for future investigation. The greatest advantage, however, lies in the opportunity of having a new space in which to seek solutions to the practical problems encountered in applying IRT. It is not unusual to find that a problem that seems intractable in one form suddenly becomes straightforward after the appropriate transformation. Levine has provided a number of examples of problems that can be solved in the canonical space. It is probably safe to predict that more lie ahead.

REFERENCES

Andersen, E. B., & Madsen, I. Estimating the parameters of a latent population distribution. Psychometrika, 1977, 42, 357-374.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes problem). *Psychometrika*, 1969, *34*, 259-299.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, 1960.

Sanathanan, L., & Blumenthal, S. The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, 1978, *73*, 794-798.

SAMPLING VARIANCES AND COVARIANCES OF PARAMETER ESTIMATES IN ITEM RESPONSE THEORY

FREDERIC M. LORD AND MARILYN S. WINGERSKY
EDUCATIONAL TESTING SERVICE

In item response theory (IRT) the observations are in the form of an n × N matrix, with one row for each item and one column for each examinee. The joint frequency distribution of the observations depends on a vector of N "ability" parameters (one for each person) and on a matrix of item parameters. In this paper only the 3-parameter logistic model for dichotomously scored items will be considered, so there will be three item parameters (*a*, *b*, and *c*) for each of *n* items. A method will be developed for computing the *asymptotic* sampling variance-covariance matrix when both abilities and item parameters are unknown. Until this is accomplished, the standard errors of the parameter estimates are unknown, handicapping development of a goodness-of-fit test and other statistics required in applications of IRT.

If the item (or ability) parameters are known, the estimated ability (or item) parameters have independent sampling distributions. It can be shown (see Bradley & Gart, 1962) that the maximum likelihood estimates of the ability (or item) parameters are consistent. Hence, the asymptotic sampling variance for an estimated ability parameter is given by the usual formula

$$\mathrm{Var}(\hat{\tau}_r \,|\, \underset{\sim}{a}, \underset{\sim}{b}, \underset{\sim}{c}) = [\mathcal{E}(\partial \ell / \partial \tau_r)^2]^{-1} , \qquad [1]$$

where

$\hat{\tau}_r$ is the estimated ability parameter;
ℓ is the log of the likelihood; and
$\underset{\sim}{a}$, $\underset{\sim}{b}$, and $\underset{\sim}{c}$ are the known vectors of item parameters.

Similarly, the asymptotic sampling variance-covariance matrix of the estimated item parameters for an item is given by

$$\left\| \mathrm{Cov}(\hat{\tau}_v, \hat{\tau}_w \,|\, \underset{\sim}{\theta}) \right\| = \left\| \mathcal{E}\left(\frac{\partial \ell}{\partial \tau_v} \frac{\partial \ell}{\partial \tau_w} \right) \right\|^{-1} \qquad (v,w = 1,2,3) \qquad [2]$$

where $\{\hat{\tau}_v\}$ is a vector consisting of the estimated *a*, *b*, and *c* for a single item and $\underset{\sim}{\theta}$ is the known vector of abilities. The right-hand side is the inverse of a 3 × 3 matrix.

When neither item nor ability parameters are known, all parameters are often estimated simultaneously by maximum likelihood. In the (Rasch) case where there is only one parameter per item, Haberman (1977) has shown that all parameter estimates will converge to their true values (i.e., will be consistent) when the number of examinees and the number of test items become large simultaneously. Empirical results suggest that consistency probably also holds when all parameters are estimated simultaneously under the 3-parameter model. If so, it is reasonable that the asymptotic sampling variance-covariance matrix of all estimated parameters will be given by the usual formula

$$\| \mathrm{Cov}(\hat{\tau}_p, \hat{\tau}_q) \| = \| \mathcal{E} \frac{\partial \ell}{\partial \tau_p} \frac{\partial \ell}{\partial \tau_q} \|^{-1} \qquad (p,q = 1,2, \ldots, M) , \qquad [3]$$

where $M = 3n + N - 2$
and $\underset{\sim}{\tau} \equiv \{\tau_p\} \equiv \{a_1, b_1, c_1, a_2, b_2, c_2, \ldots, a_n, b_n, c_n; \theta_1, \theta_2, \ldots, \theta_{N-2}\}'$.

Since standard errors are urgently needed in practical work where all parameters are estimated simultaneously by maximum likelihood, this report compares numerical values provided by Equation 3 with values provided by Equations 1 and 2 and with empirically observed sampling fluctuations. The comparisons to be presented suggest that Equation 3 provides useful values for the desired standard errors.

There are several special problems that arise in the evaluation and practical utilization of Equation 3, problems that do not arise in the situation where Equations 1 and 2 are appropriate:

1. Until an origin and scale are specified, the parameters are not identifiable.
2. The mathematical formulation is complicated by the choice of origin and scale.
3. The usual choice of origin and scale when estimating IRT parameters is inconvenient for mathematical purposes.
4. The numerical values of the sampling variances are very much affected by the choice of origin and scale.
5. Equation 3 requires the inversion of a matrix of order N + 3n - 2, where N may be several thousand.

These problems will be considered in subsequent sections.

1. Parameterization

The appropriate likelihood function is (Lord, 1980)

$$L(\underset{\sim}{a}, \underset{\sim}{b}, \underset{\sim}{c}; \underset{\sim}{\theta} | U) = \prod_{i=1}^{n} \prod_{a=1}^{N} P_{ia}^{u_{ia}} Q_{ia}^{1-u_{ia}} \qquad [4]$$

where $\underset{\sim}{\theta}$ is the vector of the N ability parameters; $\underset{\sim}{a}$, $\underset{\sim}{b}$, and $\underset{\sim}{c}$ are each a vector of *n* item parameters; $U \equiv \| u_{ia} \|$ is the matrix of item responses u_{ia} (= 0 or 1); finally, $Q_{ia} \equiv 1 - P_{ia}$; and P_{ia} is the item response function, the probability of a correct answer by examinee *a* to item *i*. Each given P_{ia} is a function of θ_a and of a_i, b_i, and c_i, but not of any other parameters. In numerical work here, P_{ia} will be considered to be the 3-parameter logistic function

$$P_{ia} \equiv c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta_a - b_i)]} \; . \qquad [5]$$

For mathematical purposes, however, it is only necessary to state that P_{ia} is an increasing function of θ_a.

If some constant is added to all θ_a and the same constant is subtracted from all b_i, all P_{ia} will be unchanged. This means that the origin used for measuring ability is entirely arbitrary. If each θ_a and each b_i is multiplied by some constant and each a_i is divided by the same constant, again, all P_{ia} will be unchanged. This means that the unit used to measure ability is entirely arbitrary. Since the origin and unit of the θ_a can be changed without changing Equation 4, it follows that $\underset{\sim}{\theta}$, $\underset{\sim}{a}$, $\underset{\sim}{b}$, and $\underset{\sim}{c}$ are not identifiable and cannot be estimated from Equation 4 without further specification.

To conform to a commonly used procedure, the origin and scale could be chosen so that for some specified group of examinees the mean of the θ_a is 0 and the variance is 1. This is not convenient mathematically, however. Instead, two other methods of specifying the origin and scale will be used, even though this will complicate matters later on when the results are applied in practice. In the first method, without loss of generality, arbitrary numerical values will be assigned to θ_{N-1} and to θ_N.

The $M \equiv N + 3n - 2$ likelihood equations are

$$0 = \sum_{i=1}^{n} \sum_{a=1}^{N} (u_{ia} - P_{ia}) \frac{P_p^{ia}}{P_{ia}Q_{ia}} \qquad (p = 1,2, \ldots, M) \qquad [6]$$

where $P_p^{ia} \equiv \partial P_{ia}/\partial \tau_p$.

2. Fisher Information Matrix

The Fisher information matrix on the right of Equation 3 now has as a typical element

$$I_{pq} \equiv \mathcal{E}\left(\frac{\partial \ell}{\partial \tau_p} \frac{\partial \ell}{\partial \tau_q}\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{a=1}^{N} \sum_{b=1}^{N} \frac{P_p^{ia} P_q^{jb}}{P_{ia}Q_{ia}P_{jb}Q_{jb}} \operatorname{Cov}(u_{ia}, u_{jb})$$

$$(p,q = 1,2, \ldots, M) \; . \qquad [7]$$

Because of local independence and random sampling of examinees,

$$\operatorname{Cov}(u_{ia}, u_{jb}) = \delta_{ij}\delta_{ab}P_{ia}Q_{ia} \qquad [8]$$

where $\delta_{st} = 1$ if $s = t$, $\delta_{st} = 0$, otherwise. Thus, the typical element is

$$I_{pq} \equiv \sum_{i=1}^{n} \sum_{a=1}^{N} \frac{P_p^{ia} P_q^{ia}}{P_{ia} Q_{ia}} \qquad (p,q = 1,2, \ldots, M) . \qquad [9]$$

Note that P_p^{ia} is zero unless either p and a refer to the same person or p and i refer to the same item. Thus,

$$\| I_{pq} \| = \left[\begin{array}{cccc|cccc} \underset{\sim}{S}_1 & 0 & \cdots & 0 & \underset{\sim}{f}_{11} & \underset{\sim}{f}_{12} & \cdots & \underset{\sim}{f}_{1N'} \\ 0 & \underset{\sim}{S}_2 & \cdots & 0 & \underset{\sim}{f}_{21} & \underset{\sim}{f}_{22} & \cdots & \underset{\sim}{f}_{2N'} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \underset{\sim}{S}_n & \underset{\sim}{f}_{n1} & \underset{\sim}{f}_{n2} & \cdots & \underset{\sim}{f}_{nN'} \\ \hline \underset{\sim}{f}'_{11} & \underset{\sim}{f}'_{21} & \cdots & \underset{\sim}{f}'_{n1} & t_1 & 0 & \cdots & 0 \\ \underset{\sim}{f}'_{12} & \underset{\sim}{f}'_{22} & \cdots & \underset{\sim}{f}'_{n2} & 0 & t_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \underset{\sim}{f}'_{N'1} & \underset{\sim}{f}'_{N'2} & \cdots & \underset{\sim}{f}'_{N'n} & 0 & 0 & \cdots & t_{N'} \end{array} \right] , \qquad [10]$$

where

$N' \equiv N - 2$;

$\underset{\sim}{S}_i$ is the 3 × 3 Fisher information matrix for a_i, b_i, and c_i;

t_a is the Fisher information for examinee a; and

$\underset{\sim}{f}_{ia}$ is the 3 × 1 joint Fisher information vector for item i and examinee a:

$$\underset{\sim}{f}_{ia} \equiv \frac{\partial P_{ia} / \partial \theta_a}{P_{ia} Q_{ia}} \begin{bmatrix} \partial P_{ia} / \partial a_i \\ \partial P_{ia} / \partial b_i \\ \partial P_{ia} / \partial c_i \end{bmatrix} . \qquad [11]$$

3. Matrix Inversion

The following general formula for inverting a partitioned matrix may be applied to Equation 10:

$$\left[\begin{array}{c|c} \underset{\sim}{S} & \underset{\sim}{F} \\ \hline \underset{\sim}{F}' & \underset{\sim}{T} \end{array} \right]^{-1} \equiv \left[\begin{array}{c|c} \underset{\sim}{S}^{-1} + \underset{\sim}{S}^{-1} \underset{\sim}{F} \underset{\sim}{Z}^{-1} \underset{\sim}{F}' \underset{\sim}{S}^{-1} & -\underset{\sim}{S}^{-1} \underset{\sim}{F} \underset{\sim}{Z}^{-1} \\ \hline -\underset{\sim}{Z}^{-1} \underset{\sim}{F}' \underset{\sim}{S}^{-1} & \underset{\sim}{Z}^{-1} \end{array} \right] \qquad [12]$$

where

$$Z \equiv T - F'S^{-1}F \ . \tag{13}$$

The matrix S is easily inverted, since it is a diagonal supermatrix:

$$S^{-1} = \left\| {}_{o}\; S_i^{-1}\; {}^{o} \right\| \ . \tag{14}$$

The notation on the right denotes a diagonal matrix with diagonal elements S_i^{-1}. These last are easily computed, since each S_i is only a 3 × 3 matrix.

All the matrix operations indicated on the right side of Equation 12 can be carried out on the computer without difficulty with one exception: the inversion of Z, which is $N' \times N'$. The approximation used here to invert Z relies on grouping the θ_a into 16 class intervals of width .5, covering the range $-5 \leq \theta_a \leq 3$. Each θ_a in a given class interval is replaced by the midpoint of the interval.

Now, T will be a diagonal supermatrix $T \equiv \left\| {}_{o}\; T_g\; {}^{o} \right\|$, where $T_g \equiv t_g I$ is a scalar matrix with dimensions $N_g \times N_g$, and N_g is the number of people in class interval g. Also, F will be a row vector of 16 matrices, all the columns of any one matrix being identical:

$$F \equiv \{f_1 1_1', \ f_2 1_2', \ \ldots, \ f_{16} 1_{16}'\} \ , \tag{15}$$

where $f_g = \{f_{ia}\}$ for any examinee a in class interval g, and 1_g is a unit vector whose length is N_g.

The product $F'S^{-1}F$ can now be written as a 16 × 16 supermatrix:

$$F'S^{-1}F = \left\| 1_g f_g' S^{-1} f_h 1_h' \right\| \ . \tag{16}$$

Denote the scalar $f_g' S^{-1} f_h$ by w_{gh}. This now gives

$$Z \equiv T - \left\| M_{gh} \right\| \ , \tag{17}$$

$$M_{gh} \equiv w_{gh} 1_g 1_h' \ . \tag{18}$$

For computation purposes, Z still has N' rows and columns, not just 16. For the usual sample size, it is still not feasible to invert Z with a standard inversion program.

Consider the problem of inverting Z_{11}, the $N_1 \times N_1$ upper left corner of Z. By Equation 17, Equation 18, and a standard formula,

$$Z_{11}^{-1} \equiv [T_1 - w_{11} 1_1 1_1']^{-1} \equiv T_1^{-1} + \frac{w_{11} T_1^{-1} 1_1 1_1' T_1^{-1}}{1 - w_{11} 1_1' T_1^{-1} 1_1} \ . \tag{19}$$

Since $\mathbf{T}_1 \equiv t_1\mathbf{I}$, where t_1 is scalar, this becomes

$$\mathbf{Z}_{11}^{-1} = \frac{\mathbf{I}}{t_1} + \frac{w_{11}\mathbf{1}_1\mathbf{1}_1'}{t_1^2 - t_1 w_{11} N_1} . \qquad [20]$$

Next, the upper left 2 × 2 supermatrix in $\mathbf{Z}$ can be inverted as in Equation 12, using the standard formula for the inversion of a partitioned matrix:

$$\begin{bmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22} \end{bmatrix}^{-1} = \left[\begin{array}{c|c} \mathbf{Z}_{11}^{-1} + \mathbf{Z}_{11}^{-1}\mathbf{Z}_{12}\mathbf{H}^{-1}\mathbf{Z}_{21}\mathbf{Z}_{11}^{-1} & -\mathbf{Z}_{11}^{-1}\mathbf{Z}_{12}\mathbf{H}^{-1} \\ \hline -\mathbf{H}^{-1}\mathbf{Z}_{21}\mathbf{Z}_{11}^{-1} & \mathbf{H}^{-1} \end{array}\right] , \qquad [21]$$

where $\mathbf{H} = \mathbf{Z}_{22} - \mathbf{Z}_{21}\mathbf{Z}_{11}^{-1}\mathbf{Z}_{12}$.
It can be seen that $\mathbf{H}$ has the same general form as $\mathbf{Z}_{11}$ and can thus be inverted as in Equation 19; so Equation 21 can readily be calculated.

Next, substitute Equation 21 for $\mathbf{Z}_{11}^{-1}$ in the foregoing procedure and repeat this procedure in such a way as to invert the upper left 3 × 3 supermatrix in $\mathbf{Z}$. A total of 15 repetitions inverts the 16 × 16 supermatrix $\mathbf{Z}$. Equation 12 is now used for one final inversion, the result being the desired variance-covariance matrix of all $N + 3n - 2$ parameters.

The 16 × 16 variance-covariance supermatrix for the $\hat{\theta}_a$ consists of 256 blocks. The elements are all the same within a block except for diagonal blocks, each of which has a variance (instead of a covariance) repeated along its diagonal. Any two examinees in the same class interval will have identical Var $\hat{\theta}$ and identical sampling covariances with any other given parameter estimate.

4. Reparameterization

In Section 1 of this paper, in order to have identifiable parameters, an origin and scale were chosen so that θ_{N-1} and θ_N had arbitrary preassigned values. Any other choice of origin and scale would result in a linear transformation of parameters. The likelihood function would remain unchanged for every pattern of item responses.

The choice of unit (but not the choice of origin) has one completely obvious effect on the sampling errors of parameter estimates. If the unit is changed, the standard errors for the $\hat{b}$'s and $\hat{\theta}$'s will be multiplied by the ratio of the new scale unit to the old scale unit. The standard errors for the $\hat{a}$'s will be divided by this ratio. A second important effect is easily overlooked: The standard error of the maximum likelihood estimator depends not only on the choice of scale but also on how the (origin and) scale is specified.

Suppose that the true numerical values of all θ_a ($a = 1, \ldots, N$) are specified on some arbitrary scale. Suppose next that the test is too difficult for examinee N. This means that the likelihood function is rather insensitive to variations in θ_N. If testing could be repeated with several parallel test forms, a wide range of estimates of θ_N would be found. In such a situation, the difference between true θ_{N-1} and θ_N clearly cannot be estimated well from the examinee responses. If the scale is defined by treating θ_N and θ_{N-1} as known, the estimates of every θ_a may fluctuate grossly, simply because the scale unit $\theta_N - \theta_{N-1}$ is not well determined by the data.

Suppose next that all examinees are relabeled so that examinees N - 1 and N are not the same examinees as before. The ability scale has not been changed; it is the procedure for defining the scale that has been changed. The true θ for each examinee is still the same as before. Suppose the new examinees N - 1 and N are both at ability levels where the test measures accurately. If, further, the true θ_{N-1} and θ_N are substantially different from each other, the difficulty of the previous paragraph disappears: Throughout the ability range where the test is designed to measure accurately, the standard errors of all $\hat{\theta}_a$ may be reasonably small.

For example, suppose on some scale $\theta_1 = -3$, $\theta_2 = -2$, $\theta_3 = -1$, $\theta_4 = 0$, $\theta_5 = 1$, $\theta_6 = 2$, $\theta_7 = 3$. This same scale can be specified in terms of any two of these θ's. The standard errors that are obtained will depend in an overwhelming way not just on the ability scale but on how it is specified. The standard errors cannot be rectified by some simple procedure, such as multiplying each by a constant.

For this reason, the procedure for specifying the ability scale should depend only on parameters or functions of parameters that are accurately determined by the data. A robust mean of the θ_a might seem attractive; however, any function of the θ_a is counterindicated by the fact that sometimes $\hat{\theta}_a = \pm \infty$.

The procedure used here is to choose a set of m discriminating, moderately easy items and a set of r discriminating, moderately difficult items. Hereafter, the origin and unit for the new parameters, to be denoted by capital letters, will be defined so that the mean of the (true) B-parameters for the easy items is 0, and the mean for the difficult items is 1.

The new parameters are related to the old parameters (from either Section 2 or from Section 5) by linear transformations:

$$A_i \equiv ka_i\ ,\quad B_i \equiv K + b_i/k\ ,\quad C_i \equiv c_i\ ,\quad \Theta_a \equiv K + \theta_a/k\ ,$$

$$(a = 1,2,\ \ldots,\ N;\ i = 1,2,\ \ldots,\ n)\ , \qquad [22]$$

where k and K are transformation constants to be determined. Since

$$\bar{B}_0 \equiv \frac{1}{m}\sum^{m} B_i = 0\ ,\quad \bar{B}_1 \equiv \frac{1}{r}\sum^{r} B_i = 1\ , \qquad [23]$$

the values of k and K are found by substituting Equation 22 into Equation 23 and solving for k and K:

$$k = \bar{b}_1 - \bar{b}_0 \; , \quad K = -\bar{b}_0/k \; , \tag{24}$$

where $\bar{b}_0$ and $\bar{b}_1$ are means for m and r items, respectively.

To find the variance-covariance matrix for estimates of the upper-case parameters, rewrite Equation 22 as

$$\Theta_a = (\theta_a - \bar{b}_0)/k \; , \quad A_i = ka_i \; , \quad B_i = (b_i - \bar{b}_0)/k \; , \quad C_i = c_i \; . \tag{25}$$

Because of the special properties of maximum likelihood estimators, Equations 25 still hold when estimators are substituted for parameters. Thus, the sampling variances and covariances for estimates of the new parameters can be computed from the sampling variances and covariances already obtained at the end of Section 3. Formulas for doing this can be written from Equation 25 by using the "delta" method (Kendall & Stuart, 1969, chap. 10). For example,

$$\left.\begin{aligned}
\mathrm{Cov}(\hat{A}_i,\hat{\Theta}_a) &= \mathrm{Cov}(\hat{a}_i,\hat{\theta}_a) - \mathrm{Cov}(\hat{a}_i,\hat{\bar{b}}_0) - \frac{\theta_a - \bar{b}_0}{k}\,\mathrm{Cov}(\hat{a}_i,\hat{k}) \\
&\quad + \frac{a_i}{k}\,\mathrm{Cov}(\hat{\theta}_a,\hat{k}) - \frac{a_i}{k}\,\mathrm{Cov}(\hat{\bar{b}}_0,\hat{k}) - \frac{a_i(\theta_a - \bar{b}_0)}{k^2}\,\mathrm{Var}\,\hat{k} \; , \\
\mathrm{Cov}(\hat{\bar{b}}_0,\hat{k}) &= \mathrm{Cov}(\hat{\bar{b}}_1,\hat{\bar{b}}_0) - \mathrm{Var}\,\hat{\bar{b}}_0 \; , \\
\mathrm{Cov}(\hat{\bar{b}}_1,\hat{\bar{b}}_0) &= \frac{1}{mr}\sum^{m}\sum^{r}\mathrm{Cov}(\hat{b}_i,\hat{b}_j) \; .
\end{aligned}\right\} \tag{26}$$

5. Parameter Estimation

The maximum likelihood estimators (MLE) satisfy the likelihood Equations 6. In Equation 6 there is one equation for each parameter omitting θ_{N-1} and θ_N. If all $N + 3n \equiv M + 2$ MLE are linearly transformed as, for example, in Equation 22, the transformed parameters will still satisfy the likelihood equations.

Since the origin and scale for the new parameters are chosen to satisfy Equation 23, then the appropriate k and K are obtained from Equation 24 after replacing $\bar{b}_0$ and $\bar{b}_1$ by their MLE. The likelihood function, Equation 4, is unaffected by these linear transformations.

The computer program LOGIST identifies the parameters by still another choice of origin and scale:

1. A certain truncated mean of the $\hat{\theta}_a$ ($a = 1, 2, \ldots, N$) is set equal to zero.
2. A certain truncated standard deviation of the $\hat{\theta}_a$ is set equal to one.

The usual lower-case symbols for parameters will be used for parameters on this LOGIST scale. This should not cause confusion, since the lower-case parameters of the first three sections will not be needed again.

Beginning with LOGIST $\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$, and $\hat{\theta}_a$ and determining $\underline{k}$ and K so that $\hat{\bar{B}}_0 = 0$ and $\hat{\bar{B}}_1 = 1$, then the $\hat{A}_i$, $\hat{B}_i$, $\hat{C}_i$ $(i = 1, 2, \ldots, n)$ and the $\hat{\theta}_a$ $(a = 1, 2, \ldots, N)$, calculated by substituting estimated values into Equation 22, will still satisfy the likelihood equations. The upper-case parameter estimates so obtained do not strictly have the sampling variance-covariance matrix found theoretically at the end of Section 4 because this matrix does not take into account the scaling used during the LOGIST parameter estimation. However, the empirically determined variance-covariance matrix of MLEs and the corresponding theoretical matrix should agree approximately if the two methods (Section 4 and LOGIST methods) of choosing the origin and scale are about equally effective. The remaining task is to compare these two methods.

6. Recapitulation

At different points, three different arbitrary scales have been used for parameters:

1. θ_N and θ_{N-1} are assigned arbitrarily.
2. The origin is set at $\bar{B}_0$, the unit is $\bar{B}_1$.
3. The origin is set at a truncated mean of the θ_a, the unit is a truncated standard deviation of the θ_a.

Scale 1 (denoted by lower-case symbols) is most convenient mathematically for the difficult task of inverting the M × M information matrix. Scale 1 is not useful for practical purposes, however, since its use grossly inflates all the sampling variances.

Scale 2 (denoted by upper-case symbols) seems the simplest choice in an attempt to keep the sampling error in the estimated origin and unit as small as possible. The sampling variances computed for Scale 1 are transformed (see Equation 26) to values appropriate for Scale 2. Although Scale 2 is not the familiar one, the two item sets used to specify the scale can be chosen so that the numerical values of $\hat{A}_i$, $\hat{B}_i$, and $\hat{C}_i$ differ little from the familiar $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ produced by LOGIST.

Scale 3 (hereafter denoted by lower-case symbols) is the scale used by LOGIST.

7. Empirical Estimation Procedures

As already stated, the theoretical results can be trusted only if they are shown to be in reasonable agreement with empirical results. For this purpose, artificial data $\|u_{ia}\|$ were created representing the administration of a 45-item test to a random sample of 1,500 examinees. The 1,500 θ_a were a spaced sample drawn from a distribution of abilities from a regular test administration. Six replicate matrices of $\|u_{ia}\|$ were independently generated, using the same item parameters and the same 1,500 θ_a. The variation in responses across these matrices thus represents random fluctuations in u_{ia} for fixed a_i, b_i, c_i, and θ_a.

Further replication was also built in: Items 16-30 and Items 31-45 had the same item parameters as Items 1-15. The true lower-case and upper-case item parameters are shown in Table 1 for Items 1-15.

Table 1
True (Upper Case) Item Parameters

Item	A	a	B	b	C or c
1	.96	.99	-1.75	-2.01	.17
2	.34	.35	-1.33	-1.61	.17
3	1.34	1.38	-.80	-1.09	.17
4	.76	.78	-.48	-.77	.17
5	.41	.42	-.38	-.67	.17
6	.90	.92	-.04	-.34	.17
7	.90	.92	.16	-.15	.17
8	1.04	1.06	.31	.00	.17
9	1.31	1.34	.42	.11	.13
10	1.46	1.50	.58	.26	.34
11	.85	.87	.79	.46	.17
12	.60	.62	.90	.57	.17
13	1.06	1.09	1.01	.68	.25
14	1.36	1.39	1.23	.90	.29
15	1.46	1.50	1.50	1.16	.18

Six independent runs were made on LOGIST, one for each group of 1,500 examinees. For each run separately, $\hat{\bar{b}}_0$ was calculated from Items 4-9, 19-24, 34-39; $\bar{b}_1$ was calculated from Items 10-15, 25-30, and 40-45. It is convenient for the ultimate interpretation of the standard errors to be obtained that the true $\bar{b}_1 - \bar{b}_0 = .671 - (-.305) = .976$. Since this is close to 1.0, the scale unit for the capitalized parameters is very close to the scale unit for the lower-case (LOGIST) parameters.

For each run separately, all lower-case parameter estimates were linearly transformed as in Equation 22 to the upper-case scale, using estimated $\underline{k}$ and K values. For the data reported in subsequent sections, the true $k = .9\overline{76}$ and the true $K = .312$. Since the six runs are independent, an unbiased empirical estimate of the sampling variance of any parameter estimate $\hat{T}$ is given by

$$s^2_{\hat{T}} \equiv \frac{6}{5}\left[\frac{1}{6}\sum^{6} \hat{T}^2 - \left(\frac{1}{6}\sum^{6} \hat{T}\right)^2\right], \qquad [27]$$

the sum being across the six LOGIST runs. If the $\hat{T}$ in Equation 27 were normally distributed, $s^2_{\hat{T}}/\sigma^2_{\hat{T}}$ would have an F distribution with 5 and ∞ degrees of freedom.

Since three different items have identical item parameters, the $s^2_{\hat{T}}$ for a single item parameter can be averaged across these three items to yield the best available unbiased estimate:

$$\bar{s}^2_T \equiv \frac{1}{3} \sum^{3} s^2_T . \qquad [28]$$

Note that it would be incorrect to pool all 18 values of $\hat{T}$ in an equation like Equation 27, since $\hat{T}$s from the same LOGIST run are not independent.

If T_i and S_i represent two different item parameters in the same item,

$$\bar{s}(\hat{T}_i,\hat{S}_i) \equiv \frac{1}{3} \sum^{3} s(\hat{T}_i,\hat{S}_i) \qquad [29]$$

which is the same as Equation 28 except that covariances are substituted for variances. If $\hat{T}_i$ and $\hat{S}_j$ represent item parameters in different items, then there are nine different sample covariances to be summed:

$$\bar{s}(\hat{T}_i,\hat{S}_j) = \frac{1}{9} \sum^{3}\sum^{3} s(\hat{T}_i,\hat{S}_j) . \qquad [30]$$

If T is an ability parameter, Equation 27 still holds. For purposes of this paper, T can be replaced by Θ and the equation written

$$\bar{s}^2_{\hat{\Theta}} = \frac{1}{N_g} \sum^{N_g} s^2_{\hat{\Theta}} \qquad [31]$$

where the sum is over all examinees in group *g*. When Θ is at the midpoint of interval *g*, this average should be roughly equal to the $\sigma_{\hat{\Theta}}$ obtained in Section 4.

If subscripts *a* and *b* denote different examinees in group *g*,

$$\bar{s}(\hat{\Theta}_a,\hat{\Theta}_b) = \frac{2}{N_g(N_g - 1)} \sum_{a>b}\sum s(\hat{\Theta}_a,\hat{\Theta}_b) , \qquad [32]$$

where the sum is over all pairs of examinees in group *g*. If *a* and *b* denote examinees in groups *g* and *h*, respectively ($g \neq h$), then

$$\bar{s}(\hat{\Theta}_a,\hat{\Theta}_b) = \frac{1}{N_g N_h} \sum_{a=1}^{N_g} \sum_{b=1}^{N_h} s(\hat{\Theta}_a,\hat{\Theta}_b) . \qquad [33]$$

Finally, if T_i is an item parameter and examinee *a* is in group *g*, then

$$\bar{s}(\hat{T}_i,\hat{\Theta}_a) = \frac{1}{3N_g} \sum^{3} \sum^{N_g} s(\hat{T}_i,\hat{\Theta}_a) . \qquad [34]$$

In computing Equations 31 through 34, examinees are grouped on their true values, not on their estimated values.

A problem arises when an examinee obtains a perfect score or a zero score. In this case his/her $\hat{\Theta}$ is infinite and cannot be advantageously used. Instead of making some ad hoc adjustment, the 17 examinees for whom this occurred were simply removed from the group of examinees studied, leaving N = 1,483. This has the effect of slightly biasing $\bar{s}_{\hat{\Theta}}$ for the remaining most extreme Θ values.

8. Numerical Standard Errors

Since the c parameter of an easy item usually cannot be accurately estimated, LOGIST in ordinary use does not estimate them individually. This would prevent the empirical standard errors of Section 7 from agreeing with the theoretical standard errors of Section 4. The main purpose of this paper, however, is to show that the method of Section 4 can give useful results; thus, the empirical and theoretical standard errors reported here are all estimated or calculated under the condition that the true values of c_i are known for i = 1, 2, 3, 4, 5, 12. Items 1 through 5 are easy items, and Item 12 was included because of its low a_i. For empirical work, the true c values were supplied to LOGIST, which held them fixed while estimating all other parameters. For theoretical work, the rows and columns of Equation 10 corresponding to c_1, c_2, c_3, c_4, c_5, and c_{12} were simply deleted from the information matrix, Equation 10, before inversion.

Table 2 compares the empirical standard errors of Section 7 for $\hat{B}$ with the theoretical standard errors of Section 4. The last three columns show the squared ratios for the three replications of each item; each of these ratios will have an F distribution with 5 and ∞ degrees of freedom, provided (1) $\hat{B}$ has a normal sampling distribution, (2) $\hat{B}$ is unbiased, and (3) the theoretical $\sigma_{\hat{B}}$ from Section 4 is the same as the $\sigma_{\hat{B}}$ for the LOGIST estimates. An F above 2.21 or below .229 is significant at the (two-tailed) 10% level. Eleven of the ratios are significant. The number of ratios less than one is approximately the same as the number of ratios greater than one.

In the past, the only available standard errors for item parameters assumed that the θ were known. Such standard errors for $\hat{B}$, for known θ, are given in the second column of the table. A comparison of the second and third columns shows very close agreement except for the three easiest items (1,2,3). For these three items, the new theoretical value is larger and agrees better with the empirical value. This gives support to the new theoretical values. That the empirical values (from Section 7) tend to be larger than the theoretical (from Section 4) could be due to n and N not being large enough for asymptotic results. A second likely explanation is that the LOGIST choice of origin and scale does not yield as accurate estimates as the choice in Section 4.

Table 3 makes comparisons for $\hat{A}$. Again, the standard errors of $\hat{A}$ with θ unknown agree closely with the results when θ is known. The empirical standard errors, although correlating well with the theoretical, seem to be larger. Eleven of the F ratios are significant. Similar statements apply to Table 4, which shows the comparisons for $\hat{C}$.

Table 2
Theoretical and Empirical Standard Errors for $\hat{B}$

Item	$\sigma_{\hat{B}\mid\theta}$ (θ known)	$\sigma_{\hat{B}}$ (Sect. 4)	$\bar{s}_{\hat{B}}$ (Sect. 7)	$s^2_{\hat{B}} / \sigma^2_{\hat{B}}$ Rep. 1	Rep. 2	Rep. 3
1*	.110	.156	.183	.23	.56	3.34†
2*	.186	.201	.237	1.76	1.49	.93
3*	.045	.071	.063	1.38	.59	.41
4*	.060	.068	.066	.90	.76	1.17
5*	.100	.099	.103	.37	.40	2.48†
6	.125	.121	.131	.28	.63	2.63†
7	.113	.110	.100	1.24	.65	.58
8	.084	.083	.088	2.31†	.97	.16†
9	.055	.055	.067	.37	2.63†	1.47
10	.069	.069	.106	3.19†	3.62†	.33
11	.100	.097	.122	1.45	2.55†	.70
12*	.094	.091	.087	.85	1.27	.66
13	.086	.083	.094	1.01	1.20	1.57
14	.077	.076	.111	1.19	1.49	3.75†
15	.072	.075	.093	.40	2.62†	1.65

†Significant at 10% level.
*The C parameter for these items was treated as known.

Table 3
Theoretical and Empirical Standard Errors for $\hat{A}$

Item	$\sigma_{\hat{A}\mid\theta}$	$\sigma_{\hat{A}}$	$\bar{s}_{\hat{A}}$	$s^2_{\hat{A}}/\sigma^2_{\hat{A}}$ Rep. 1	Rep. 2	Rep. 3
1*	.088	.105	.141	.95	.91	3.60†
2*	.044	.046	.039	.88	.51	.74
3*	.097	.117	.094	1.39	.32	.22†
4*	.060	.065	.080	.89	2.77†	.86
5*	.045	.047	.054	.63	2.44†	.93
6	.103	.102	.123	1.54	.30	2.51†
7	.105	.105	.147	1.30	2.25†	2.35†
8	.113	.115	.159	1.29	3.20†	1.29
9	.123	.128	.182	1.89	3.39†	.80
10	.184	.193	.160	.71	.55	.79
11	.115	.120	.132	1.42	1.85	.34
12*	.060	.060	.076	.95	2.94†	.94
13	.151	.157	.187	2.40†	1.08	.79
14	.209	.218	.240	1.32	.91	1.43
15	.222	.233	.182	.25	.65	.93

†Significant at 10% level.
*The C parameter for these items was treated as known.

Table 4
Theoretical and Empirical Standard Errors for $\hat{C}$

Item*	$\sigma_{\hat{C}\mid\theta}$	$\sigma_{\hat{C}}$	$\bar{s}_{\hat{C}}$	$s^2_{\hat{C}}/\sigma^2_{\hat{C}}$ Rep. 1	Rep. 2	Rep. 3
6	.056	.058	.063	.39	.44	2.79†
7	.049	.050	.038	.40	.35	.95
8	.037	.037	.045	3.08†	.76	.43
9	.024	.025	.039	.80	4.71†	1.83
10	.025	.026	.034	2.24†	2.68†	.27
11	.036	.037	.043	.98	2.67†	.41
13	.026	.027	.037	.89	1.88	2.90†
14	.019	.020	.028	2.98†	2.55†	.43
15	.015	.015	.016	.64	1.23	1.71

†Significant at 10% level.
*C_1, ..., C_5, and C_{12} were treated as known.

Table 5 compares standard errors for $\hat{\theta}$. Column 3 will be discussed later. Columns 4 and 5 show standard errors of $\hat{\theta}$ corresponding to the θ value in the first column; column 6, however, is computed from Equation 3 for the group of N_g people falling in the class interval with midpoint θ. There is good agreement between empirical and theoretical standard errors except for $\theta < -1.5$. For low θ, asymptotic results do not appear with the usual n and N.

Table 5
Theoretical and Empirical Standard Errors for $\hat{\theta}$

θ	N_g	All C_i Unknown $\sigma_{\hat{\theta}}$	C_1 to C_5 and C_{12} Treated as Known $\sigma_{\hat{\theta}\mid \underset{\sim}{A}, \underset{\sim}{B}, \underset{\sim}{C}}$	$\sigma_{\hat{\theta}}$	$\bar{s}_{\hat{\theta}}$
-2.75	10	2.090	.951	.966	*
-2.25	35	1.296	.686	.699	1.134
-1.75	93	.861	.516	.525	.797
-1.25	219	.607	.400	.404	.427
-.75	332	.456	.341	.342	.332
-.25	326	.349	.295	.295	.279
.25	227	.278	.262	.263	.274
.75	136	.261	.260	.261	.286
1.25	77	.303	.289	.290	.349
1.75	25	.422	.384	.387	.412
2.25	3	.628	.575	.580	*
2.75	0	.931	.874	.878	*

*Not computed because of small N_g.

Table 5 shows close agreement of the standard error from Sections 2 through 4 with the standard error of $\hat{\theta}$ when the item parameters are known. The agreement shown here and in previous tables suggests that Equation 1 is a good approximation to the diagonal of Equation 3 and, similarly for item parameters, that Equation 3 agrees well with the empirical standard errors.

A comparison of the third and fifth columns in Table 5 shows what happens to $\sigma_{\hat{\theta}}$ when all C_i must be estimated from the data. For $\theta < -1$, $\sigma_{\hat{\theta}}$ is sharply affected; for $0 < \theta < 2.5$, there is very little effect.

Table 6 contains the squared ratios of the empirical standard errors to the theoretical standard errors for the five θ values closest to the midpont of the intervals, and within at least .1 of the midpoint. Two of the groups had only two abilities within this restriction. If similar caveats apply as for the item parameters, these ratios will have an F distribution with 5 and ∞ degrees of freedom. Only 8 of the ratios are significant at the two-tailed 10% level, and only 16 are greater than 1.

Table 6
F Ratios for $\hat{\theta}$

θ	$s^2_{\hat{\theta}}/\sigma^2_{\hat{\theta}}$				
-2.75	3.73†	4.41†			
-2.25	.85	.78	.43	11.34†	1.16
-1.75	.57	1.90	1.62	.32	18.95†
-1.25	.98	.63	.96	.95	.77
-.75	.26	.94	.63	.81	.63
-.25	.71	1.81	.73	.04†	.48
.25	.18†	.98	.74	.80	.77
.75	.61	.35	1.41	1.21	.64
1.25	2.76†	1.82	.98	1.08	1.84
1.75	.67	.41	1.08	1.45	1.78
2.25	.11†	.36			
2.75*					

†Significant at 10% level.
*There was no θ between 2.65 and 2.85.

Table 7 presents the theoretical standard errors of $\hat{A}$, $\hat{B}$, and $\hat{C}$, obtained by the method of Sections 2 through 4, when all C_i must be estimated from the data. It is interesting to compare these values with those in Tables 2 through 4 where $C_1, \ldots, C_5$, and C_{12} were treated as known. The standard errors of $\hat{B}_1$ to $\hat{B}_5$ are increased drastically by ignorance of C_1 to C_5; all other $\sigma(\hat{B}_i)$ are much increased, except for $i = 11$, 13, and 14. All $\hat{A}_i$ show sharply increased standard errors. For items for which C_i must be estimated, on the other hand, the standard errors of $\hat{C}_i$ are little affected by knowledge or ignorance of $C_1, \ldots, C_5, C_{12}$. A likely explanation for this is that errors in estimating the scale unit B_1 affect the standard errors of the $\hat{A}_i$ and $\hat{B}_i$, but not of the $\hat{C}_i$.

Table 7
Standard Errors from Equation 3 of Item
Parameters When All C_i Must Be Estimated

Item	$\sigma_{\hat{B}}$	$\sigma_{\hat{A}}$	$\sigma_{\hat{C}}$
1	.52	.23	.60
2	2.54	.13	.72
3	.35	.32	.10
4	.26	.15	.14
5	.97	.10	.32
6	.19	.18	.07
7	.16	.18	.06
8	.14	.21	.041
9	.12	.26	.026
10	.11	.32	.026
11	.10	.18	.039
12	.18	.14	.07
13	.09	.23	.027
14	.08	.31	.020
15	.10	.33	.015

Tables 2 through 7 present some illustrative answers to the question, how do estimation errors on one set of items affect the accuracy of estimated parameters for a different set of items? Such effects could not be quantified until now, since the standard error of an item parameter estimate was previously known only for fixed $\underset{\sim}{\theta}$. It is only through the sampling fluctuations of $\hat{\underset{\sim}{\theta}}$ that estimation errors for one item can affect parameter estimates for another item.

With 18 C_i treated as known, the Fisher information matrix inverted for this study has $3 \times 45 - 18 + 1{,}498 = 1{,}615$ rows and columns. The matrix inversion by the method of Section 4 used 1232K bytes of memory on an IBM 3031 and took 32 seconds. The computer program dealt with a 45-item test; it did not take advantage of the fact that the 45 items consisted of three replicate sets of 15 items each.

In order to verify the numerical accuracy of the inversion, the information matrix and the variance-covariance matrix were multiplied. The result was an identity matrix accurate to 10 decimal places. The variance-covariance matrix obtained in double precision agreed with the matrix obtained in quadruple precision to all six decimal places printed.

9. Sampling Covariances and Correlations

When item parameters are known, $\hat{\theta}_a$ and $\hat{\theta}_b$ $(a \neq b)$ are uncorrelated. When ability parameters are known, estimated item parameters for different items are uncorrelated. When both item and ability parameters are estimated, in general, all estimates are correlated. The computer printout of the sampling correlations for the present study consisted of 10 correlation matrices. These need only be summarized here.

Table 8 shows the theoretical (T) and empirical (E) correlations between estimates of two different parameters for the same item. The correlations are generally substantial. For comparison, the theoretical correlations when the abilities are known are included. The empirical correlations are obtained by dividing the estimated sampling covariance by the square roots of the estimated sampling variances. If the empirical correlations here have roughly 15 degrees of freedom, their standard error is roughly $(1 - \rho^2)/\sqrt{15} = .26(1 - \rho^2)$. In view of their standard errors, there is very satisfactory agreement of empirical with theoretical correlations.

Table 8
Theoretical (T) and Empirical (E) Sampling Correlations Between 2-Parameter Estimates for the Same Item

		$\rho_{\hat{A}\hat{B}}$			$\rho_{\hat{B}\hat{C}}$			$\rho_{\hat{A}\hat{C}}$	
Item	$\rho_{\hat{A}\hat{B}\mid\theta}$	T	E	$\rho_{\hat{B}\hat{C}\mid\theta}$	T	E	$\rho_{\hat{A}\hat{C}\mid\theta}$	T	E
1	.82	.86	.87						
2	.80	.82	.88						
3	.55	.70	.65						
4	.42	.52	.76						
5	.35	.38	.44						
6	.73	.70	.53	.92	.90	.92	.76	.70	.53
7	.67	.64	.66	.90	.88	.77	.76	.71	.79
8	.56	.52	.26	.83	.81	.85	.72	.67	.58
9	.37	.33	.50	.69	.67	.87	.65	.60	.81
10	.41	.42	.68	.69	.68	.93	.61	.61	.74
11	.40	.42	.70	.75	.74	.89	.77	.77	.83
12	-.55	-.51	-.79						
13	.22	.21	.06	.60	.59	.66	.69	.70	.67
14	.06	.03	.35	.45	.42	.61	.58	.59	.68
15	-.19	-.25	-.81	.25	.21	-.18	.53	.54	.56

Table 9 shows both theoretical and empirical correlations for the $\hat{B}_i$ ($i = 1, 2, \ldots, 15$). The corresponding standard errors are given in parentheses in the diagonal. The only theoretical correlations above .20 are among $\hat{B}_1$, $\hat{B}_2$, $\hat{B}_3$, and $\hat{B}_4$. These are the four easiest items. Any error in estimating the scale unit $\bar{B}_1 - \bar{B}_0$ would seriously affect all these items in the same way. It is difficult to draw other useful generalizations from this table.

The corresponding table for the $\hat{A}_i$ ($i = 1, 2, \ldots, 15$) shows only three theoretical correlations above .20: $\rho_{13} = .27$, $\rho_{14} = .20$, $\rho_{34} = .23$. With two exceptions ($\rho_{67} = -.013$, $\rho_{6,12} = -.002$), all theoretical correlations are positive. The highest theoretical correlation among the $\hat{C}_i$ ($i = 6, 7, \ldots, 11$ and $13, 14, 15$) is $\rho_{67} = .04$. All correlations are positive.

Table 9

Experimental (E) and Theoretical (T) Standard Errors (Diagonals) and Correlations for Transformed B

Item		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	E	(183)	141	284	045	193	-158	-028	-004	014	-122	289	016	-031	-064	-147
	T	(156)	264	509	334	158	-124	-128	-088	-040	034	-001	039	-005	-022	-050
2	E	141	(237)	541	286	-092	-036	126	-005	-252	-105	029	-064	-247	360	-040
	T	264	(201)	284	184	078	-066	-069	-046	-017	026	007	022	-002	-018	-039
3	E	284	541	(063)	308	040	-091	056	004	-274	-279	268	007	-298	348	-155
	T	509	284	(071)	377	151	-131	-139	-093	-032	048	008	044	-008	-032	-068
4	E	045	286	308	(066)	120	-072	-179	038	-362	-308	-007	192	-443	343	218
	T	334	184	377	(068)	066	-130	-130	-089	-040	029	003	028	-005	-018	-039
5	E	193	-092	040	120	(103)	-228	-126	-072	046	-205	236	086	-051	-193	126
	T	158	078	151	066	(099)	-117	-113	-088	-062	-004	-009	013	-003	001	002
6	E	-158	-036	-091	-072	-228	(131)	014	076	-041	107	-153	002	016	122	-085
	T	-124	-066	-131	-130	-117	(121)	-062	-053	-051	-016	004	-005	001	004	011
7	E	-028	126	056	-179	-126	014	(100)	-120	098	121	-050	-221	156	-018	000
	T	-128	-069	-139	-130	-113	-062	(110)	-042	-036	-011	002	-009	002	005	011
8	E	-004	-005	004	038	-072	076	-120	(088)	-068	025	-015	101	-062	081	-137
	T	-088	-046	-093	-089	-088	-053	-042	(083)	-007	004	001	-010	002	002	003
9	E	014	-252	-274	-362	046	-041	098	-068	(067)	198	037	-129	332	-357	-063
	T	-040	-017	-032	-040	-062	-051	-036	-007	(055)	023	-003	-013	002	000	-005
10	E	-122	-105	-279	-308	-205	107	121	025	198	(106)	-193	-137	270	-151	-098
	T	034	026	048	029	-004	-016	-011	004	023	(069)	-035	-052	-043	-062	-087
11	E	289	029	268	-007	236	-153	-050	-015	037	-193	(122)	041	-011	-103	-182
	T	-001	007	008	003	-009	004	002	001	-003	-035	(097)	-071	-067	-086	-107
12	E	016	-064	007	192	086	002	-221	101	-129	-137	041	(087)	-176	078	005
	T	039	022	044	028	013	-005	-009	-010	-013	-052	-071	(091)	-069	-068	-065
13	E	-031	-247	-298	-443	-051	016	156	-062	332	270	-011	-176	(094)	-341	-112
	T	-005	-002	-008	-005	-003	001	002	002	002	-043	-067	-069	(083)	-057	-060
14	E	-064	360	348	343	-193	122	-018	081	-357	-151	-103	078	-341	(111)	006
	T	-022	-018	-032	-018	001	004	005	002	000	-062	-086	-068	-057	(076)	-004
15	E	-147	-040	-155	218	126	-085	000	-137	-063	-098	-182	005	-112	006	(093)
	T	-050	-039	-068	-039	002	011	011	003	-005	-087	-107	-065	-060	-004	(075)

Note. Decimal points omitted.

The theoretical correlations between $\hat{A}_i$ and $\hat{B}_j$ ($i \neq j$) are all below .20 in absolute value, with the exception of Items 1 through 4, which vary from .14 to .38. For $\hat{B}_i$ and $\hat{C}_j$ (i j; j 1, 2, ..., 5, 12), there are no correlations above .25 in absolute value. For $\hat{A}_i$ and $\hat{C}_j$, there are no correlations above .20 in absolute value.

The theoretical correlations between $\hat{\theta}_a$ and $\hat{\theta}_b$ ($a \neq b$) are all less than .04 in absolute value. Between $\hat{\theta}_a$ and $\hat{B}_i$, the largest correlation in absolute value is .15 (when $i = 1$ and $\theta = -2.25$). Between $\hat{\theta}_a$ and $\hat{A}_i$, the largest is .12 (when $i = 1$ and $\theta = -2.25$). Between $\hat{\theta}_a$ and $\hat{C}_i$, the largest is 06.

Summary

When both abilities and item parameters are unknown, the asymptotic sampling variance-covariance matrix developed in this paper appears to provide useful values for the standard errors needed for further research in item response theory. The magnitude of the numerical values in the matrix were very much affected by the method used to define the scale. For a set of artificial data, this variance-covariance matrix compared satisfactorially with empirical results and with the variance-covariance matrices found by the usual formulas for the case where the abilities are known or where the item parameters are known.

With this matrix, the effect on other items of including items with poorly determined parameters can be studied. Including items with poorly determined $\underline{c}$'s increases the standard errors of all of the $\underline{a}$'s and $\underline{b}$'s but not of the other $\underline{c}$'s. The effect of different distributions of abilities on the accuracy of item parameters can also be studied.

The standard errors of item parameters can now be studied for a situation of common occurrence in equating and item banking: Each of two tests containing common items is administered to a different group of examinees; all parameters are estimated in the same LOGIST run. It is of particular interest to determine how the number of common items affects the standard error of the parameter estimates.

REFERENCES

Bradley, R. A., & Gart, J. J. The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, 1962, *49*, 205-214.

Haberman, S. J. Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 1977, *5*, 815-841.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 1, 3rd ed.). New York: Hafner, 1969.

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum, 1980.

ACKNOWLEDGMENTS

This work was supported in part by Contract N00014-80-C-0402, Project Designation NR 150-453, between the Office of Naval Research and Educational Testing Service. The authors are grateful to Michael Levine for pointing out an error in the original method for computing the empirical standard errors reported in Section 7.

CONFIDENCE ENVELOPES FOR ITEM RESPONSE FUNCTIONS

DAVID THISSEN
UNIVERSITY OF KANSAS

HOWARD WAINER
EDUCATIONAL TESTING SERVICE

In seminal work published in 1929, Working and Hotelling developed the formulae for the extrapolation of the concept of the confidence interval to regression lines. They provided a "confidence envelope," enclosing the population line with probability $1 - \alpha$. Such an envelope for a regression function $F(\cdot)$ has also been called a confidence "band" by Miller (1966). It is defined by the functions $L(\cdot)$ and $U(\cdot)$ illustrated in Figure 1, satisfying the relationship

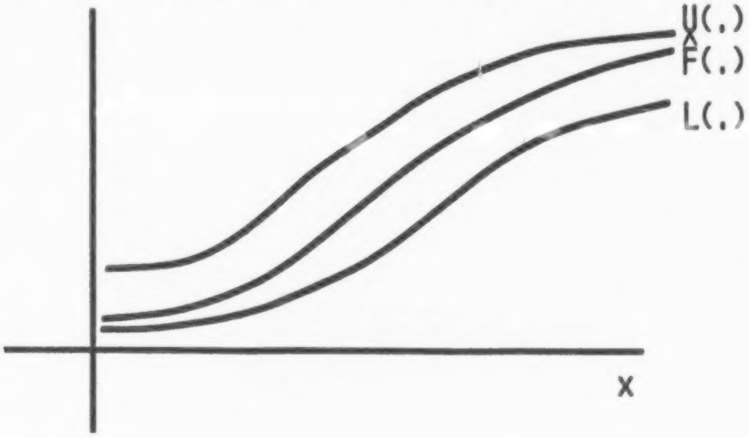
$$P\{U(\cdot) > F(\cdot) > L(\cdot), \forall x\} = 1 - \alpha \qquad [1]$$

in which $U(\cdot)$ and $L(\cdot)$ are upper and lower bounds of the envelope, respectively, and $F(\cdot)$ is the function. The envelope problem for linear models has been treated extensively by Working and Hotelling (1929), Roy (1957), Miller (1966), and others; but all use the intimate relationships of linear models with multinormal error to develop computational formulae for $U(\cdot)$ and $L(\cdot)$. This paper describes and illustrates a methodology for constructing confidence envelopes satisfying Equation 1 for any regression function $F(\cdot)$ that is monotonic in its parameters.

It is easier to understand the nonlinear problem if the basic concepts underlying the standard linear regression envelope are considered first. This requires the simultaneous consideration of two spaces in a duality diagram: the "function space" on the left of Figure 2 and the "parameter space" on the right. Although there are a number of algebraic procedures for the derivation of the hyperbolic envelope boundaries shown in Figure 2 (Miller, 1966), a geometric interpretation will be considered here.

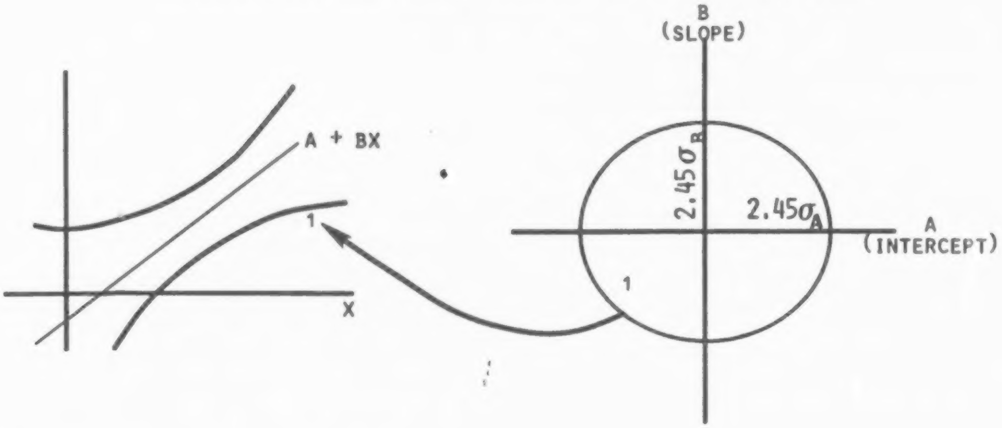
The right panel of Figure 2 illustrates the conventional sampling space for the parameters of linear regression. The circle encloses the central 95% of a bivariate normal with its mean equal to the estimates and covariance matrix given by classical maximum likelihood results. The phrase "95% confidence interval" describes the probability that the circle "covers" the true value of the parameter vector, so the circle in the right panel of Figure 2 illustrates a confidence region for the slope-intercept combination. In general, the circle in Figure 2 is an ellipse with major and minor axes of length equal to $2.45\sigma_a$ and $2.45\sigma_b$. The coefficient 2.45 equals $[\chi^2(.95)]^{1/2}$ with 2 df. See Thissen and

Figure 1
Generalized Illustration of a Confidence Envelope for a Fitted Curve $\hat{F}(\cdot)$, in Which the Envelope is Bounded Above by $U(\cdot)$ and Below by $L(\cdot)$



Wainer (1982b) for a complete explanation. For purposes of exposition, a rescaling that yields a circle is used here. This can be done without loss of generality.

Figure 2
The Confidence Envelope in Linear Regression: The Right Panel Shows the Central 95% Region in the Parameter (Slope, Intercept) Space, While the Left Panel Shows the Corresponding Hyperbolic Envelope for the Fitted Line in the Function Space



The illustration of the confidence region in the parameter space is not very informative, however. It is often more useful to map the confidence region into the function space. Each point of the parameter space corresponds to a line in the function space. The center of the parameter space corresponds to the estimated regression line, and all lines represented by slope-intercept combinations within the circle are in the envelope described by the hyperbolae in Figure 2. Thus, the hyperbolic boundaries describe a 95% envelope for the regression line because they bound lines corresponding to the central 95% confidence region in the parameter space.

Indeed, each point on the hyperbolae corresponds to a point (a set of parameters) on the boundary of the circle in the parameter space. For instance, the point marked 1 on the lower hyperbola represents a point with low slope and lower intercept--marked 1 in the parameter space--so that it gives the lowest value of $a + b_{x_1}$ for an a and b in the central 95% region of the parameter space. The 95% envelope pictured in Figure 2 is thus simply a representation in the function space of the circle in the parameter space. The lower hyperbola in Figure 2 could be plotted by plotting, for each value x_0, the values

$$y_0 = a_{x_0} + b_{x_0} x_0 \qquad [2]$$

where (a_{x_0}, b_{x_0}) is the point on the circle that minimizes y. There is a closed form solution for the linear case, so numerical minimization is not required. The principle, however, generalizes readily to nonlinear regression; explicit minimization is required with some models to find the envelope boundary.

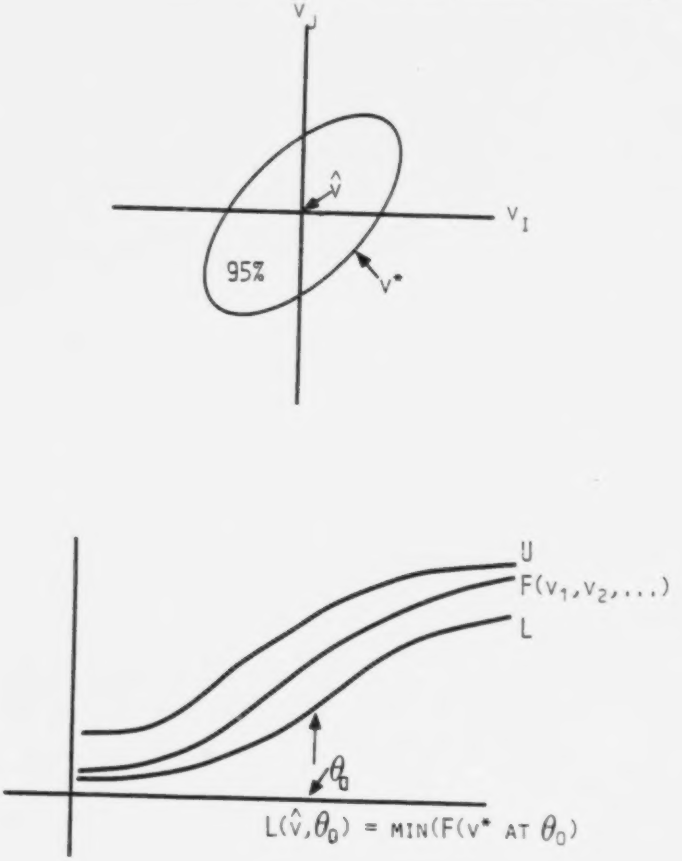
Logistic Item Response Models

The authors (Thissen & Wainer, 1982b) have previously described a computational technique for plotting the boundaries for the regression equivalents of 2- and 3-parameter logistic functions. The problem is described schematically in Figure 3. Any item response model may be denoted $F(\theta;v)$ in which θ is the latent ability and v represents the parameters of the nonlinear regression model; in the more usual notation, $v = [a, b]$ for the 2-parameter logistic model and $v = [a, b, c]$ for the 3-parameter model. The central 95% confidence region of the sampling distribution for the estimates of the parameter vector v is illustrated by the bivariate normal ellipse in the top panel of Figure 3. Error covariance between the parameters is suggested in the ellipse because the errors of estimate of the parameters of item response functions are correlated in most circumstances; the single exception is when $E(\theta) = b$ (Thissen & Wainer, 1982a). The nonlinear envelope problem may be stated in the same way as in the linear case: To plot the lower boundary $L(\theta;\hat{v})$ for parameter estimates $\hat{v}$, find for each value θ_0 the minimum value of $F(\theta_0;v^*)$ where v^* is constrained to lie on the boundary of the 95% confidence region in the parameter space.

There are a number of ways to solve such a constrained minimization problem. The authors' solution (Thissen & Wainer, 1982b) was to reparameterize the problem to become a minimization problem in terms of the angle defining each vector v^*. Once the minimization problem is solved, the resulting points are plotted as $L(\theta;\hat{v})$. The negative of the same function is minimized to give $U(\theta;\hat{v})$.

Figure 4 is an illustration of the resulting 95% envelope for a 2-parameter logistic of the form

Figure 3
The Confidence Envelope Generalized to Nonlinear Regression:
the Upper Panel Shows the Central 95% Region in the
Parameter Space for Parameters with Correlated Error)
and the Lower Panel Shows the Corresponding Envelope


V_J
$\hat{v}$
V_I
95%
v^*
U
$F(v_1, v_2, \ldots)$
L
θ_0
$L(\hat{v}, \theta_0) = \mathrm{MIN}(F(v^* \text{ AT } \theta_0)$

$$F(\theta;[a,b]) = 1 \;/\; \{1 + \exp[-a(\theta - b)]\} \qquad [3]$$

with parameters [1,0] and a sample of size 100 drawn from a standard normal distribution. It graphically suggests the wide variety of item characteristic curves that might fit the data under these circumstances. Figure 5 shows the much narrower bounds and a steeper slope given by a sample size of 500. The envelope in Figure 5 is not symmetrical because the location is -1, and that is not the mean of the population distribution, which is zero. Thissen and Wainer (1982) have shown that if the location of the item is not the same as the mean of the population distribution, the sampling errors of a and b are correlated, and that covariance translates into asymmetry in the regression envelope. By contrast, the envelope for Figure 4 is symmetrical, since b = 0 there.

Figure 4
A 95% Confidence Envelope for a 2-Parameter Logistic ICC, with a = 1, b = 0, and Sample Size 100

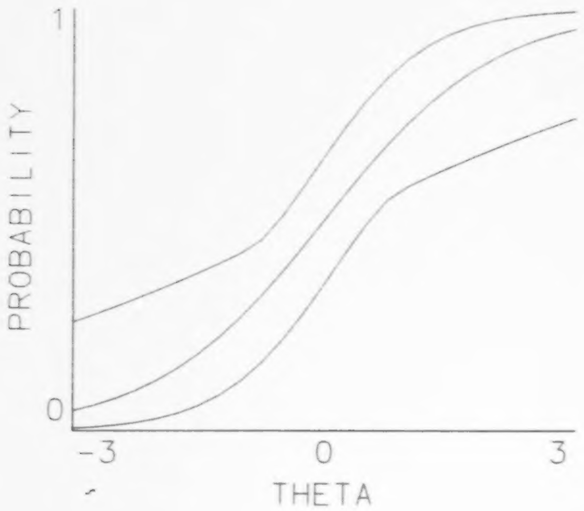


Figure 5
A 95% Confidence Envelope for a 2-Parameter Logistic ICC, with a = 1.5, b = -1, and Sample Size 500.

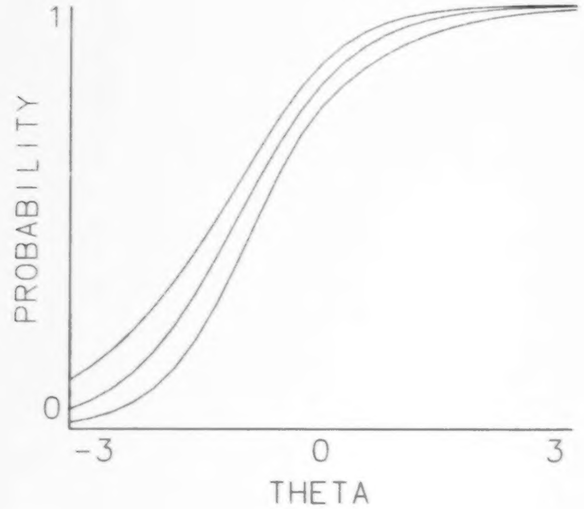
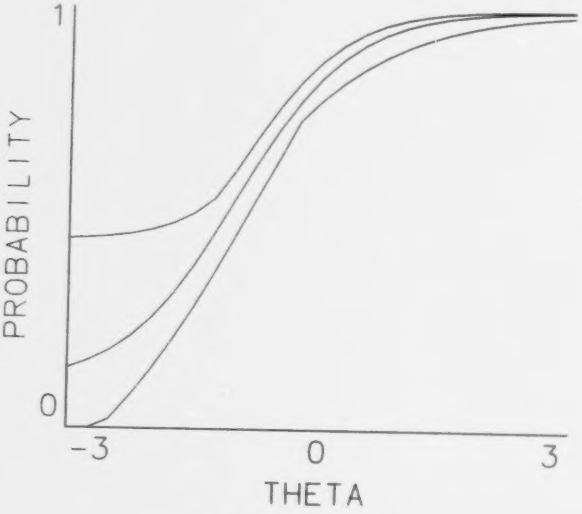


Figure 6 illustrates the envelope for the 3-parameter logistic,

$$F(\theta;[a,b,c]) = c + (1 - c) / \{1 + \exp[-a(\theta - b)]\} \qquad [4]$$

Although such plots can be made, the process is made difficult by local minima in the minimization problem described above. Lord and Stocking (1982) have shown that the envelope problem has local minima for the 3-parameter logistic model, so great care is required in the computations to assure finding the true minimum to plot. Nevertheless, such plots are potentially useful tools. Note that the envelopes constructed will be too narrow (i.e., the estimates will look optimistically good) if the solution is a local rather than a global minimum.

Figure 6
A 95% Confidence Envelope for a 3-Parameter Logistic ICC,
with $a = 1.5$, $b = -1$, $c = 0.1$, and Sample Size 1,000



N-Line Plots

A Bayesian interpretation would denote the sampling distribution of the parameters in a regression problem as the posterior density of the parameters obtained from a suitable uninformative prior distribution. In the Bayesian view, parameter vectors near the estimates are more likely (have greater posterior density) than parameters more distant from the posterior mean, although all may lie within the central 95% region. While plots of the $(1 - \alpha)$ envelope define precisely a $(1 - \alpha)$ confidence interval for the fitted curve, they do not graphically display the density of the curves within the envelope. Indeed, for a frequentist statistician, there is no "density of curves within the envelope"; but the Bayesian formulation derives such a concept from the posterior density. N-line plots are plots which (approximately) fill the $(1 - \alpha)$ envelope with curves, in proportions roughly corresponding to their posterior likelihood. N-line plots consist of plots of N lines drawn using parameters randomly sampled from the posterior distribution for the parameters.

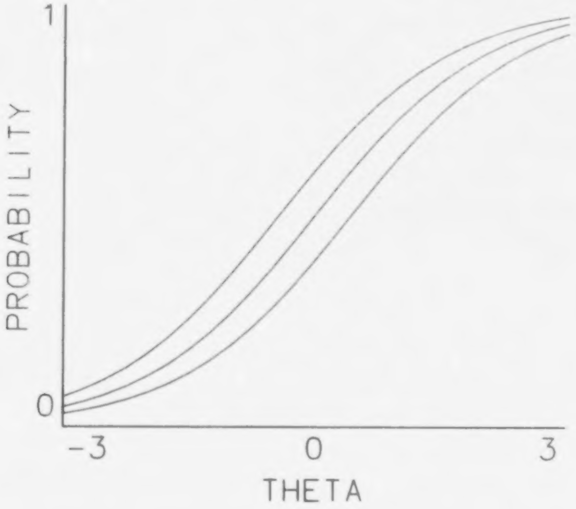
A 1-Parameter Example: 25-Line Plots

The simplest widely used item response model is the 1-parameter logistic model, in which (for many practical purposes) the slope may be taken to be a known constant; this is similar to a regression model of the form

$$F(\theta;b) = 1 / \{1 + \exp[-(\theta - b)]\} \qquad [5]$$

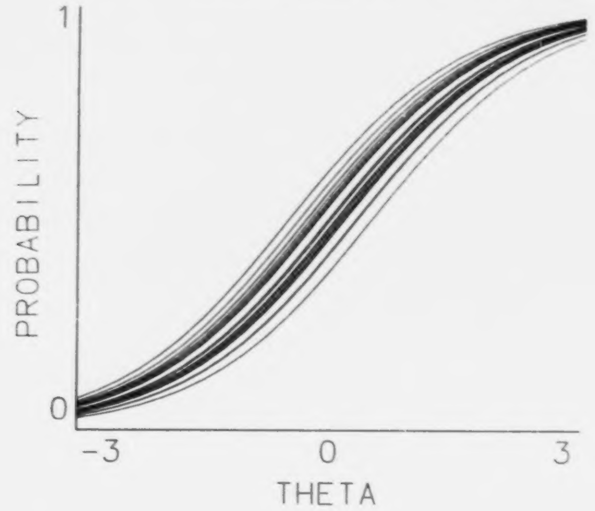
in which b is a location parameter and the only estimated parameter in the model. The computation of the $(1 - \alpha)$ envelope for the fitted curve is trivial: If $\hat{b}$ is the estimate of b, with a standard error of σ_b, then the upper boundary for the 95% envelope $U(\cdot)$ is $F(\theta;\hat{b} - 1.96\sigma_b)$ and the lower boundary for the envelope is $L(\cdot) = F(\theta;\hat{b} + 1.96\sigma_b)$. A typical 95% envelope for such a function is plotted in Figure 7, with the modal (fitted) curve in the center of the envelope. The boundaries of the envelope plot describe the central 95% confidence interval for the fitted curve.

Figure 7
A 95% Confidence Envelope for a 1-Parameter Logistic ICC, with b = 0 and Sample Size 100



In Figure 8, following the description of the posterior density for b as $N(\hat{b},\sigma_b^2)$, the curves have been plotted corresponding to 25 random deviates from a normal density with mean $\hat{b}$ and standard deviation σ_b. Note that such randomly sampled curves are dense near the curve corresponding to $\hat{b}$ and, albeit more thinly toward the edges, fill the 95% envelope of Figure 7.

Figure 8
25 1-Parameter Logistic ICCs, with Locations Sampled Randomly from the Posterior Distribution for b = 0 and Sample Size 100



The envelope described by the upper and lower extreme curves of the 25-line plot is, in expectation, almost exactly equal to the true 95% envelope. This is because the distance between the upper and lower boundaries of the 95% envelope plot is determined by plotting $F(\theta;\hat{b} + 1.96\sigma_b)$ and $F(\theta;\hat{b} - 1.96\sigma_b)$ for two values of b, centered on $\hat{b}$, which have a range of $2(1.96)\sigma_b$. The random deviates which yield the 25-line plot are also centered on b, and the expected value of their range is $3.93\sigma_b$, which is very nearly $2(1.96)\sigma_b$. (Values of the expected value of the range of various size samples from normal distributions are given by Pearson and Hartley, 1966.) The boundaries of the 95% envelope and the expected values of the highest and lowest curves of the 25-line plot are thus in the same place. The 25-line plot gives a clearer graphic description of the envelope for the curve.

A 2-Parameter Example: 85-Line Plots

The regression version of the 2-parameter logistic model, in which

$$F(\theta;a,b) = 1 / \{1 + \exp[-a(\theta - b)]\} \qquad [6]$$

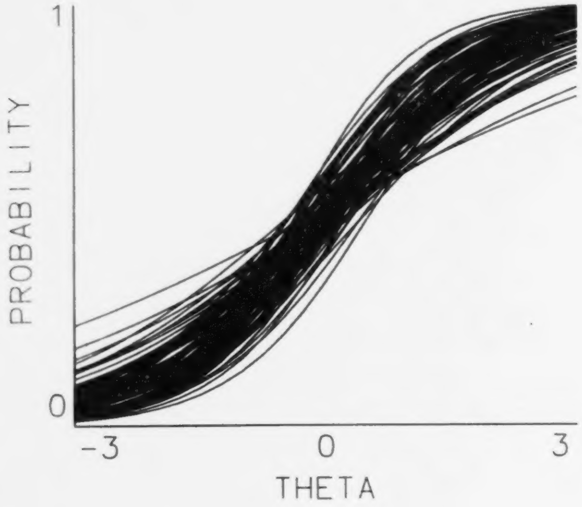
provides more interesting N-line plots, but they are more costly to make. The vector $[a,b]$ is assumed estimated with posterior density $N([\hat{a},\hat{b}],\Sigma_{ab})$. If the upper and lower extreme curves of the N-line plot are to equal (in expectation) the boundaries of the 95% envelope, the appropriate N-line plot should have as its (squared) range on any line through the origin of the density (shown in Figure 3) Mahalanobis distance $2\chi^2_{(1-\alpha)}$. That is, on any line through the origin of the 2-dimensional space in Figure 3, the expected range of the projections of the random deviates of the N-line plots should be the same as the distance between the two boundaries of the central confidence region on that line. Every value of θ has as its upper and lower $(1 - \alpha)$ envelope boundaries the parameters at the intersections of the $(1 - \alpha)$ confidence region boundary with some line through the origin of the [a,b] space. Therefore, the requirement that the expected value of the range of the projections of the random sample on any such line be equal to $(1 - \alpha)$ confidence region boundary difference gives the desired result: The expected location of the upper and lower N-line curves are along the upper and lower $(1 - \alpha)$ envelope boundaries.

For the 2-dimensional case, N (regrettably) is 85. This is determined by noting that the [a,b] space of Figure 3 is a projection of the standard 2-dimensional multinormal, and the central 95% region of the standard bivariate normal has radius equal to the square root of the value of

$$\left[\chi_{(95\%)}\right]^{1/2}(2) = 2.45 \; . \qquad [7]$$

The expected value of the standard normal deviates sampled must be twice this (4.9) to provide an N-line plot of the required breadth. Pearson and Hartley's (1966) table indicates that the required sample size is 85. Consequently, N-line plots can be made from the two-space by sampling 85 standard normal deviates, transforming them to have the [a,b] density in Figure 3 and plotting the resulting curves. Figure 9 shows an illustrative 85-line plot; the corresponding 95% envelope plot was in Figure 4. The 85-line plot shows clearly how the odd shape of the 95% envelope plot arises from the distribution of curves. Flat curves make up U(•) on the left and and L(•) on the right, while steep curves make up L(•) on the left and U(•) on the right.

Figure 9
85 2-Parameter Logistic ICCs, with Parameters Sampled from the Posterior Distribution for a = 1, b = 0, and Sample Size 100



BEST COPY AVAILABLE

Figure 10
25 2-Parameter Logistic ICCs, with Parameters Sampled from the Posterior Distribution for a = 1, b = 0, and Sample Size 100

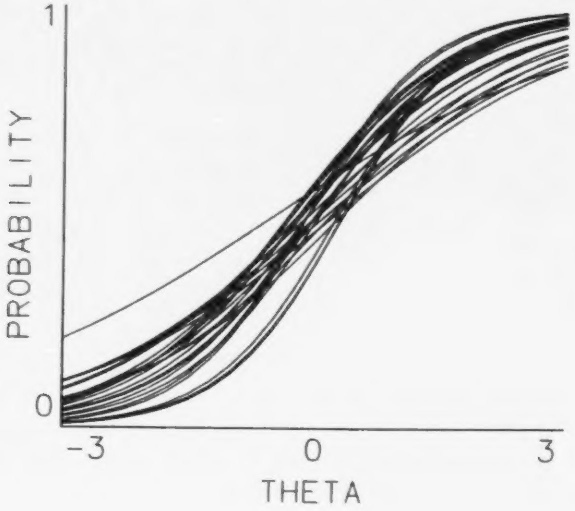
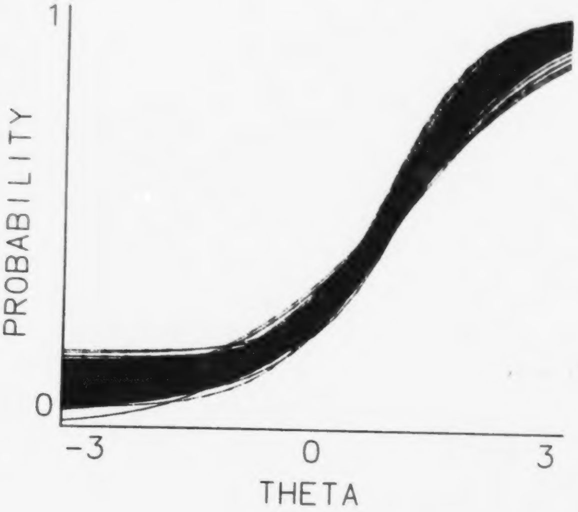


Figure 11
235 3-Parameter Logistic ICCs, with Parameters Sampled from the Posterior Distribution for a = 1.5, b = 1, c = .1, and Sample Size 1,000

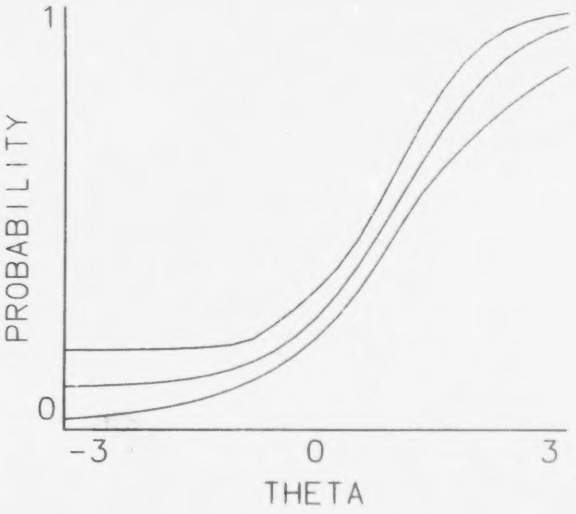


Eighty-five is a large number of lines to plot. Twenty-five-line plots are equivalent (in expectation) to only 85% envelopes but, as Figure 10 shows, they give much of the detail available in 85-line plots for a third of the cost.

The 3-Parameter Model

Through similar arguments to those of the preceding sections, N-line plots for 3-parameter models required N = 235 to match (in expectation) 95% envelope plots. Figure 11 shows a 235-line plot for the 3-parameter logistic model described above, and Figure 12 is the 95% envelope plot. The plotting of 235 lines requires an inordinate amount of time. The width of a 25-line plot for 3-parameter models is equivalent to a 72% envelope, so the much simpler plots could be useful in many applications.

Figure 12
A 95% Confidence Envelope for the 3-Parameter ICC of Figure 11



Conclusions

A description of the sampling variation of a fitted function in terms of the standard errors or the variance of the posterior density of the parameters may not be the most useful specification of that variation. Functions are usually fitted for the sake of the functions, not the parameters. This is especially true of item response functions when they are to be concatenated to form posterior density functions for ability associated with certain observed response vectors.

Confidence envelopes provide a description of the sampling variation of item response curves in the space of the fitted functions. They can be used to give the data analyst a clear idea of the class of item response curves which are compatible with the data. N-line plots may be used to show the width of the envelope, as well as the slopes and relative posterior density of the included curves. These techniques can be quite useful in applied item analysis.

REFERENCES

Lord, F. M., & Stocking, M. Personal communication, 1982.

Miller, Jr., R. G. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

Pearson, E. S., & Hartley, H. O. Biometrika Tables for Statisticians, Vol. 1 (3rd ed.). London: Cambridge University Press, 1966.

Roy, S. N. Some aspects of multivariate analysis. New York: John Wiley & Sons, 1957.

Thissen, D., & Wainer, H. Some standard errors in item response theory. Psychometrika 47, 397-412.

Thissen, D., & Wainer, H. Confidence envelopes for monotonic functions: Principles, derivations, and an example (AFHRL-TR-82-37). Brooks Air Force Base TX: Air Force Human Resources, Manpower and Personnel Division, 1982.

Working, H., & Hotelling, H. Application of the theory of error to the interpretation of trends. Journal of the American Statistical Association Supplement, 1929, 24, 73-85.

ACKNOWLEDGMENTS

This work was supported, in part, by the Air Force Human Resources Laboratory under Contracts F41689-81-6-0012 to McFann-Gray Associates and F41689-82-C-0020 to the Educational Testing Service. The authors were helped toward the solution of this problem and toward clearer prose in its description through conversations and comments on earlier drafts by R. Darrell Bock, Henry Braun, Benjamin Fairbank, Paul Holland, Frederic Lord, Donald Rubin, and Glenn Shafer.

DISCUSSION

MICHAEL V. LEVINE
UNIVERSITY OF ILLINOIS

Sampling Variances and Covariances of Parameter Estimates in Item Response Theory

Knowledge of sampling distributions of estimated item parameters is important in most item response theory (IRT) applications and essential in studies of equating and item bias. Lord and Wingersky's contribution consists of three parts: (1) the identification and analysis of the complex effects of the choice of a measurement scale on sampling distributions, (2) a new technique for approximating sampling distributions, and (3) an empirical evaluation of the new technique. I will discuss only the new technique.

To see why the approximations preceding Lord and Wingersky's were faulty, common sense tells us that the precision of the estimates of an item's parameters should depend upon which other items are administered along with the item. Data from a long test with easy, moderately easy, moderately difficult, and difficult items should give a much better estimate of the difficulty of a moderately difficult item than data from a short easy test with only the one moderately difficult item. However, in the approximation Lord and Wingersky are likely to replace, the variances and covariances of estimates of an item's parameters depend only upon abilities and the target item's parameters. The Lord and Wingersky approach does not have this defect. Consequently, it can be expected to be more accurate than the earlier approach. Furthermore, the new approach seems more likely to provide useful quantitative information about the effects of test composition on estimation error.

I am uncomfortable with the Lord-Wingersky algorithm in its present form because: (1) it involves a very large number of arithmetic operations, (2) it requires assigning abilities to blocks, (3) it seems very complicated and difficult to program, and (4) it did not produce item parameter variance and covariance estimates that were clearly superior to those produced by its simplistic predecessor. In fact, more than one-third of the standard errors for item parameters obtained by the earlier method were as close or closer to the data than those obtained by the new method.

In its present form the algorithm requires the inversion of a very large matrix, the Fisher information matrix. The matrix inversion is approximated by grouping abilities into 16 more or less arbitrary blocks. The desired second moments are individual elements in the inverted matrix.

Two necessary conditions for accurate matrix inversion were checked. A doubling of arithmetic precision did not substantially change the individual entries in the obtained matrix, and the product of the obtained matrix and the information matrix was within 10 decimal places of the identity matrix.

Unfortunately, these conditions are not sufficient to guarantee that individual elements in the obtained matrix are very close to elements in the inverse of the information matrix. The same assignment to blocks was used when the arithmetic precision was increased. Thus, any inaccuracy resulting from distinguishing only 16 ability levels remains. Furthermore, the matrix product could be very close to the identity matrix, even though some of the 2.6 million individual entries were substantially in error. The 10-decimal-place accuracy of the product by itself does not guarantee high accuracy of any individual entry in the obtained matrix.

Fortunately, there is a simplification of the algorithm that eliminates the need for blocks and for inverting very large matrices. In Lord and Wingersky's notation the information matrix $\|I_{pg}\|$ has the form

$$I = \|I_{pg}\| = \begin{pmatrix} S & F \\ F' & T \end{pmatrix} \qquad [1]$$

where S is relatively small and block diagonal, and T is very large but diagonal. The upper right-hand block of I^{-1} contains the item parameter variances and covariances. Lord and Wingersky's formula for it is $S^{-1} + S^{-1}FZ^{-1}F'S^{-1}$, where Z is the very large matrix $T - F'S^{-1}F$. Since T is diagonal, it is much more easily inverted than Z. Using T^{-1} instead of Z^{-1}, the upper right-hand block can be shown to be equal to $(S - FT^{-1}F')^{-1}$.

$S - FT^{-1}F'$ is large, but much smaller than Z. For Lord and Wingersky's analysis, $S - FT^{-1}F'$ is 135 × 135, and Z is 1,500 × 1,500. At the University of Illinois, we have been working with matrices of this form obtained from two-parameter logistic models. Tim Davey uses a Cholesky decomposition to invert these matrices. We have had no problems with his program. However, the matrices in our study are 100 × 100 rather than 135 × 135.

The simplification may in fact give the same numbers as the Lord-Wingersky algorithm. However, the simplification is easier to program, does not require the assignment of examinees to blocks, appears to involve fewer arithmetic operations, and, because the matrices are smaller, facilitates the task of obtaining bounds on the error in the individual entries.

Confidence Envelopes for Item Response Functions

This paper is important because it facilitates an overdue shift in emphasis from item parameters to item response functions. All applications of IRT depend only on the item response functions. Consequently, poor item parameter estimation is sometimes inconsequential.

This seems paradoxical in view of the one-to-one correspondence between item parameters and item response functions. However, the correspondence is complicated and sometimes misleading. In Thissen and Wainer's notation, the value of an item response function at ability θ with item parameter vector $\underset{\sim}{v}$ is denoted $F(\theta;\underset{\sim}{v})$. The effect of a small change in $\underset{\sim}{v}$ will depend in a complicated way on θ, $\underset{\sim}{v}$, and the direction of the change. However, if the sum of the squared partial derivatives of F with respect to the item parameters is small, then the item characteristic curve may be well determined, even though the item parameters are not. Stated differently, the confidence interval for a point on an item response function may be very small, even though the information matrix for the estimated item parameters is nearly singular. I have seen this happen often with the three-parameter logistic model in test administrations in which one of the parameters is superfluous. This happens, for example, when the c parameter is difficult to estimate for some moderately easy items.

If all that is needed is an interval estimate for one ability level on an item characteristic curve, then the Linn, Levine, Hastings, and Wardrop (1980) method is applicable. Thissen and Wainer consider many ability levels simultaneously, however.

Thissen and Wainer present two methods for graphically showing how well a curve has been estimated. Although their methods are general, they focus on logistic item response functions with multivariate normal parameters. A difficult way to think about Thissen and Wainer's first method is in terms of their functions L and U. $L(\theta_0)$, the lower bound function evaluated at θ_0, is the minimum of the set of numbers $\{F(\theta_0, \underset{\sim}{v});\ \underset{\sim}{v}$ is in the region of highest density$\}$. $U(\theta_0)$, the upper bound function, is the maximum of the same set. The confidence envelope is the point set enclosed by the graphs of L and U.

It may be easier to think about confidence envelopes as the point set obtained by superimposing all the graphs of all the item response functions that can be obtained with parameters from the region of highest density. If the envelope is narrow, then all the considered item parameter vectors correspond to essentially the same item response function.

Thissen and Wainer's second aid to visualizing estimated curves is the "N-line plot." To generate an N-line plot, N vectors of item parameters are sampled using the multivariate distribution for the item parameters. The curves corresponding to the parameters are then superimposed. The manner in which N has been chosen seems to assure that the graphs of sampled curves will generally be between L and U and that each point on L and U will be close to a point on a sampled curve.

The N-line plots have the advantage of showing interrelations between points on item response functions at different ability levels. For example, Figure 9 provides the valuable information that very high probabilities at low ability levels imply relatively low probabilities at high levels. Confidence envelopes lack this feature. For example, in Figure 6 one suspects, but cannot verify, that all the low probabilities for $\theta = .5$ come from curves with implausibly high probabilities at $\theta = -3$.

The advantage could be retained and the "N" in N-line plots could be reduced if the parameters $\underset{\sim}{v}$ were sampled from the boundary of the region of highest density. Unfortunately, sampling from the boundary alone would not show which curves correspond to the $\underset{\sim}{v}$s with highest probability density. Sampling from the boundary, however, does seem to inexpensively approximate the confidence envelope while displaying relations between ability levels.

Reference

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. An investigation of item bias in a test of reading comprehension (Technical Report No.163). Champaign, IL: University of Illinois, Center for the Study of Reading, March, 1980.

DEVELOPMENTS IN NONPARAMETRIC ABILITY ESTIMATION

CHARLES LEWIS
UNIVERSITY OF GRONINGEN

The nonparametric approach to test theory to be discussed here has its roots in early work of Guttman, Lazarsfeld, and Meredith. References to much of the more recent work in this tradition may be found in Cliff (1979) and in Tatsuoka and Tatsuoka (in press). The most extensive treatment of the subject to date is that of Mokken (1971). Recent introductions to this work are given by Stokman and Van Schuur (1980) and by Mokken and Lewis (1982). Computer programs developed to carry out the relevant analyses are described in The STAP User's Manual (Technisch Centrum FSW, 1980). Mokken (1971) concentrated on defining, constructing, and testing unidimensional scales, based on responses to dichotomous items. Unlike most of the work in this nonparametric tradition, Mokken's emphasis is on the development of formal statistical methods to extend and support the use of descriptive indices.

The analyses to be described here assume the existence of a scale such as might result from one of Mokken's procedures. Based on their responses to items forming the scale, inferences are made regarding the abilities of individuals. In this sense, the work to be described here can be seen as an extension of Mokken's developments. Earlier references are Lewis (1981) and Mokken and Lewis (1982).

Basic Assumptions and Notation

As in classical test theory and virtually all of modern item response theory (IRT), a propensity distribution for observed score, given a particular item and person, is assumed in the following. Since attention will be restricted to dichotomous items, this must be a Bernoulli distribution and will be written as

$$\text{Prob}\ (u_{ia} = 1) = \pi_{ia} \qquad [1]$$

$$\text{Prob}\ (u_{ia} = 0) = 1 - \pi_{ia} \qquad [2]$$

for item *i*, person *a*, and observed score u_{ia}. The probability of "success" π_{ia}, which (with 0/1 scoring) is also the classical true score for u_{ia}, will serve as the basic unknown (parameter).

The only other basic assumption which will be made is that given a set of π_{ia}, the corresponding u_{ia} are conditionally independent. This allows their joint propensity distribution to be written as a product of the marginal distributions and is an assumption common to virtually all developments in modern IRT. (It is, however, stronger than the "uncorrelated errors" assumption of classical test theory.)

As developed in the next section, all additional assumptions will have to do only with the ordering of the π_{ia}. This, indeed, is the reason for using the label "nonparametric" to describe the present treatment. There will be no attempt made here to weigh the advantages and disadvantages of parametric and nonparametric approaches to IRT. The reader who is interested in the "robustness" of various approaches should, however, recall that a technique which never makes a certain assumption is automatically robust to violations of that assumption.

Alternative Ordinal Assumptions and Corresponding Definitions of Ability

Beyond the basic assumptions outlined in the last section, it will be useful to consider ordinal assumptions about the π_{ia} made in previous work, in order to motivate the assumption that will finally be used here. Since the primary emphasis of the present development is on abilities, definitions of ability that are justified by the various assumptions will also be considered.

There are two basic sorts of ordering of the π_{ia} that will be discussed. Let I be a finite set of items which are of interest for measuring a given ability, and let P be the population of persons whose abilities are of potential interest.

First, consider P by I ordering: consistent ordering of persons by items. The idea is that the π_{ia} for each item i should order the persons in the same way. Formally, for items i and j in I and persons a and b in P,

$$\pi_{ia} > \pi_{ib} \text{ implies } \pi_{ja} \geq \pi_{jb}. \qquad [3]$$

(If item i is easier for person a than for person b, then item j should be at least as easy for person a as for person b.)

Second, consider I by P ordering: consistent ordering of items by persons. This is really just a dual of the first type of ordering and states that the π_{ia} for each person a should order the items in the same way. Formally, for items i and j in I and persons a and b in P,

$$\pi_{ia} > \pi_{ja} \text{ implies } \pi_{ib} \geq \pi_{jb}. \qquad [4]$$

(If item i is easier than item j for person a, then i should be at least as easy as j for person b.)

As an illustration of these two types of ordering, a hypothetical set of π_{ia} for a Guttman scale is presented in Table 1 for four items and seven persons. Note first that due to the deterministic nature of Guttman's theory, all π_{ia} are either zero or unity. More importantly, within a Guttman scale, both types of ordering hold, and the rows and columns of Table 1 have been arranged to reflect this. Thus, for example, Person 5 will (with probability 1) give a positive response to Item 3, while Person 4 will not. P by I ordering implies that Person 5 will do at least as well as Person 4 on the remaining items. (In fact, their response probabilities are identical except for those to Item 3, so the implication holds.) For an example of the second (I by P) ordering, begin with Person 3, who can answer Item 2 but not Item 3. It is also the case that Item 2 is at least as easy as Item 3 for the other six persons, as Equation 4 requires.

Table 1
Hypothetical π_{ia} Values for Four Items
and Seven Persons, as an Illustration of
the Double Monotony Ordering Assumption
for a Guttman Scale

Items in Order	Persons in Order						
	1	2	3	4	5	6	7
1	0	1	1	1	1	1	1
2	0	0	1	1	1	1	1
3	0	0	0	0	1	1	1
4	0	0	0	0	0	0	1
Sum*	0	1	2	2	3	3	4

*Number Correct

As noted above, a Guttman scale will always satisfy both I by P and P by I ordering. This is also true for Rasch scales. Mokken (1971) used the term double monotony to identify this property in general. Thus, a set of items, I, is doubly monotone with respect to a population of persons, P, when I by P and P by I ordering both hold for all the corresponding probabilities π_{ia}. While double monotony is a highly desirable property, in the sense that it allows strong conclusions to be drawn about the persons and items for which it holds (Mokken, 1971), it is also a highly restrictive assumption that is unlikely to hold exactly for real items and persons. Thus, it is of interest to see what can be done when one or another aspect of double monotony is abandoned.

P by I Ordering Only

As a first step in this direction, suppose that persons may be ordered by items in terms of the π_{ia} (P by I ordering) but that the dual ordering assumption does not necessarily hold. Mokken (1971) referred to this property as monotone homogeneity and considered it the minimum necessary condition for asserting that a set of items measure the same (unidimensional) characteristic for a population of persons. As he showed, P by I ordering allows each person a to be assigned an ordinal latent trait value θ_a and allows nondecreasing item response functions P_i to be defined by

$$\pi_{ia} = P_i(\theta_a) .$$

A hypothetical example of monotone homogeneity is given in Table 2 (in a format similar to that of Table 1). That the persons can be ordered consistently by the items may be seen by noting that the π_{ia} values in each row of the table form a nondecreasing sequence from left to right. (To eliminate decimal points, π_{ia} is multiplied by 10 in the following tables). Noting that Person 2 finds Item 1 easier than Item 2, while the opposite holds for Person 4, establishes that the π_{ia} in Table 2 violate the I by P ordering assumption.

For convenience, θ_a in Table 2 has been identified as the sum over i of π_{ia}. This is simply the true score corresponding to the number-correct observed score for these four items. It reproduces the ordering of the persons given by the individual items, and this fact (among others) has been used by Mokken (1971) to justify the use of the number correct as an indication of ability when working with monotonely homogeneous items. Any strictly increasing transformation of the θ_a in Table 2 would, of course, be equally satisfactory as an ordinal ability measure.

Table 2
Hypothetical $\pi_{ia}(\times 10)$ Values, as an Illustration of P by I Ordering (Monotone Homogeneity)

	Persons in Order						
Items	1	2	3	4	5	6	7
1	1	5	6	6	8	9	9
2	1	4	6	7	7	9	9
3	1	2	3	4	8	8	9
4	1	1	4	4	4	4	5
θ_a(Sum)	.4	1.2	1.9	2.1	2.7	3.0	3.2
θ_a^*(Number of $\pi_{ia} \geq .5$)	0	1	2	2	3	3	4

Monotone homogeneity (P by I ordering) is a property shared by most item response models, including those which describe π_{ia} as a 2- or 3-parameter logistic or normal ogive function of ability. By giving up the second aspect of double monotony (namely, I by P ordering), item response functions are allowed to "cross" one another, in which case it is no longer possible to state unambiguously that one item is at least as difficult as another for all persons.

I by P Ordering Only

Double monotony may also be relaxed in the other direction, giving up P by I ordering but maintaining I by P. This case is discussed by Lewis (1981). Here it is clear that the definition of ability based on monotone homogeneity cannot be used (since items need not order persons consistently). If, however, an item mastery level π_m is introduced, a latent number-correct score θ_a^* may be defined as the number of items i in I for which $\pi_{ia} \geq \pi_m$. This alternative definition of ability does not in itself depend on any ordering assumptions about the π_{ia}. If, however, I by P ordering holds, then

1. All persons with the same latent number-correct score θ_a^* have mastered the same items (at mastery level π_m); and
2. $\theta_a^* > \theta_b^*$ implies that the items mastered by person b form a subset of those mastered by person a.

These properties of latent number correct are a direct stochastic generalization of those which hold for observed number correct in the case of a Guttman scale. They allow θ_a^* to be used to measure ability on a labeled ordinal scale. Property 1 states that values of θ_a^* may be labeled by the item(s) that have been mastered and Property 2 gives a specific meaning to the ordering of θ_a^*, again in terms of the items that have been mastered.

As an illustration, consider the hypothetical π_{ia} values given in Table 3. The items are ordered consistently by the persons, as can be seen from the fact that each column of π_{ia} in the table gives a nonincreasing sequence from top to bottom. That the persons are not ordered consistently by the items is shown by the fact that Item 1 is easier for Person 2 than for Person 1, while the opposite is true for Item 2. A mastery level of .5 has been chosen for determining θ_a^*, whose values range from 0 for Person 1 (reflecting that all π_{i1} are less than .5) to 4 for Person 7 (all π_{i7} being greater than or equal to .5). The fact that θ_a^* is 3 for Persons 5 and 6 implies that both have mastered Items 1, 2, and 3, but not 4 (at level .5). It also implies that they have mastered all items mastered by those persons (1 through 4) for which θ_a^* is less than 3.

Models where I by P ordering is assumed (but not double monotony) are unusual within IRT. An example is the approach based on person characteristic curves, which was discussed by Lumsden (1978). In this approach it is assumed that the π_{ia} may be written as nonincreasing functions (one for each person) of a latent item difficulty parameter. The "nonincreasing" property guarantees I by P ordering. However, if two of the functions cross, then P by I ordering is violated, since an item "below" the crossing would order the two persons one way and an item above would do the reverse.

Table 3
Hypothetical π_{ia} (×10) Values, as an
Illustration of I by P Ordering

Items in Order	Persons						
	1	2	3	4	5	6	7
1	4	5	6	7	8	8	9
2	3	2	4	6	8	8	9
3	1	2	4	3	7	6	8
4	1	1	2	3	4	3	5
θ_a^*	0	1	1	2	3	3	4

Comparing θ_a and θ^*_a

Returning to the definitions of ability, θ_a and θ^*_a, Mokken and Lewis (1982) showed that they have a natural relation in the case of double monotony. Briefly summarized, θ^*_a defines a partitioning of θ_a into ordered equivalence classes:

$$\Theta_j = \{\theta_a | \theta^*_a = j\} \ , \qquad [6]$$

where $j = 0(1)n$ in the case of *n* items. (These are ordered in the sense that $i > j$ implies that every element in Θ_i exceeds every element in Θ_j.) Thus, θ^*_a could be thought of as a *discretizing* and *labeling* of θ_a, when double monotony holds for the π_{ia}.

Mokken (1982), has pointed out that the same relationship between θ_a and θ^*_a also exists in the general case of monotone homogeneity (thus, when I by P ordering need not hold). Although the items do not have an unambiguous order in this case, they may still be ordered relative to a given mastery level π_m. This may be illustrated by returning to the example of monotone homogeneity given in Table 2. If .5 is adopted as a mastery level, the items are ordered consistently with their numbering. Thus, Item 1 is the easiest, since everyone except Person 1 has mastered it, and Item 4 is the most difficult, since only Person 7 has mastered it. (That this ordering depends on the choice of mastery level may be seen by observing that if $\pi_m = .4$, Item 3 replaces Item 4 as the most difficult.) The last row of Table 2 gives values of θ^*_a based on $\pi_m = .5$, and a comparison of these values with those of θ_a in the row above illustrates the relationship between the two. Note that although some of the inequalities among the θ_a values become equalities among θ^*_a (this is what was meant by discretizing), none of the inequalities is reversed. In other words, whenever monotone homogeneity holds, then θ^*_a may be thought of as measuring the same thing as θ_a, but doing it more coarsely. On the other hand, the labeling provided by θ^*_a may be considered as a useful return for this investment. Thus, returning to Table 2, it may be inferred from θ_a that Persons 3 and 4 have higher ability than Persons 1 and 2 but lower ability than Persons 5, 6, and 7. From θ^*_a, however, it is also known that Persons 3 and 4 have mastered Items 1 and 2, but *not* Items 3 and 4 (at the .5 level).

A "Minimal" Assumption (Relative Consistency) and Analysis

With the above extended discussion as background, the ordering assumption that will actually be used in the following may be introduced. The idea is to work with an assumption that allows item labeling of the latent number-correct score θ^*_a but is otherwise as weak as possible. To achieve this goal, neither P by I nor I by P ordering is necessary. Only consistent ordering of the π_{ia} *relative to the item mastery level* π_m used to define θ^*_a need be considered. The actual assumption will be referred to as *relative consistency*.

Specifically, begin with any pair of persons in P and any pair of items in I. If each of the persons can master just one of the two items (at level π_m), the assumption is that this is the same item for both persons. Phrased this way, the persons may be thought of as consistently ordering the items relative to π_m, a generalization of I by P ordering.

This asymmetry is actually misleading. The assumption may be equivalently stated as follows. If each of the two items can be mastered (at level π_m) by just one of the two persons, then this must be the same person for both items. Thus, the items must consistently order the persons relative to π_m, generalizing P by I ordering.

To see the equivalence of these two statements of the assumption of relative consistency, consider a formal negative expression for both: There are no items i and j in I and persons a and b in P such that the following four relations simultaneously hold:

$$\begin{array}{ll} \pi_{ia} \geq \pi_m , & \pi_{ib} < \pi_m , \\ \pi_{ja} < \pi_m , & \pi_{jb} \geq \pi_m . \end{array} \qquad [7]$$

Read by rows, Equation 7 can be seen to refer to an inconsistent ordering of the persons a and b by the items i and j. Read by columns, it refers to an inconsistent ordering of the items by the persons.

For the purposes of what follows, the importance of assuming that the pattern in Equation 7 does not occur (for the π_{ia} associated with sets P and I of interest) is to guarantee that the earlier mentioned properties--same θ_a^* score, implying same items mastered, and lower θ_a^* score, implying subset of items mastered--hold, so that θ_a^* forms a labeled ordinal scale.

To illustrate this new assumption, hypothetical π_{ia} values are given in Table 4. Using a mastery level of .5, both items and persons may be consistently ordered. (This has been done in the table.) To make verification of the two properties of θ_a^* easier, a "staircase" has been inserted in the table, separating mastery from nonmastery levels of the π_{ia}.

When π_m is changed to .4, the π_{ia} associated with the first two items and the first two persons exhibit the inadmissible pattern in Equation 7 . Thus, Person 1 can only master Item 2 and Person 2 can only master Item 1 (at level .4). This shows that relative consistency does not hold for the π_{ia} in Table 3 when $\pi_m = .4$. It also shows that neither I by P nor P by I ordering holds in this example, since the presence of pattern Equation 7 for any π_m is inconsistent with both these assumptions.

Table 4
Hypothetical π_{ia} (×10) Values, as an
Illustration of Relative Consistency ($\pi_m = .5$)

Items in Order	Persons in Order						
	1	2	3	4	5	6	7
1	3	5	7	5	9	6	9
2	4	2	4	6	8	8	7
3	4	2	3	3	7	6	8
4	1	1	2	4	4	3	5
θ^*_a	0	1	1	2	3	3	4

Statistical Analysis of an Individual Response Pattern

If the responses of individual a to Items 1 through n are denoted by

$$\underset{\sim}{u}'_a = (u_{1a},\ldots,u_{na}) \qquad [8]$$

and the corresponding probabilities of positive response by

$$\underset{\sim}{\pi}'_a = (\pi_{1a},\ldots,\pi_{na}), \qquad [9]$$

then the basic assumption of conditional independence of the u_{ia} implies

$$p(\underset{\sim}{u}_a|\underset{\sim}{\pi}_a) = \prod_{i=1}^{n} \pi_{ia}^{u_{ia}} (1 - \pi_{ia})^{1-u_{ia}} . \qquad [10]$$

For reasons discussed in detail by Lewis (1981) and mentioned by Mokken and Lewis (1982), a Bayesian aproach to making inferences about θ^*_a will be adopted here. Thus, a prior distribution for $\underset{\sim}{\pi}_a$ must be specified to describe available prior knowledge. In the above mentioned references, critical use was made of I by P ordering in selecting the prior. Here, priors will be considered which are based on the relative consistency assumption. More concretely, not only will relative consistency be assumed, but it will also be assumed that the ordering of the items relative to the chosen mastery level (π_m) is known and that the items have been numbered (from 1 to n) to reflect this ordering. Thus, for items i and j in I, $i < j$ implies that for all persons a in P, if $\pi_{ja} \geq \pi_m$, then $\pi_{ia} \geq \pi_m$. (In other words, anyone who can master item j can also master item i.)

Although the π_{ia} will not, in general, have a functional dependence on θ^*_a, it may be convenient to express prior knowledge about $\underset{\sim}{\pi}_a$ conditional on θ^*_a. Thus, the joint prior may be decomposed as

$$p(\underset{\sim}{\pi}_a,\theta^*_a) = p(\underset{\sim}{\pi}_a|\theta^*_a)p(\theta^*_a). \qquad [11]$$

Relative consistency implies that the π_{ia} cannot, in general, be stochastically independent of one another. Given θ^*_a, however, such an assumption is not ruled out and is, in fact, very convenient. Thus, it will be assumed that prior knowledge is such that the conditional distribution of $\underset{\sim}{\pi}_a$ given θ^*_a may be written as

$$p(\underset{\sim}{\pi}_a|\theta^*_a) = \prod_i p(\pi_{ia}|\theta^*_a). \qquad [12]$$

This assumption will be referred to as second-order conditional independence.

Note that the limits within which the density $p(\pi_{ia}|\theta^*_a)$ may be positive vary with the value of θ^*_a. Thus, if $\theta^*_a = j$,

$$\pi_{ia} \geq \pi_m \text{ for } i \leq j \text{ and } \pi_{ia} < \pi_m \text{ for } i > j. \qquad [13]$$

In other words, if person a has a latent number-correct score of j, then the first j items must have been mastered (at level π_m) and the last n-j items not mastered.

Continuing further with the analysis, Bayes' theorem may be used with Equations 10, 11, and 12 to obtain the joint posterior distribution for $\underset{\sim}{\pi}_a$ and θ^*_a, given $\underset{\sim}{u}_a$:

$$p(\underset{\sim}{\pi}_a,\theta^*_a|\underset{\sim}{u}_a) \propto p(\theta^*_a)\prod_i\left[\pi_{ia}^{u_{ia}}(1-\pi_{ia})^{1-u_{ia}}p(\pi_{ia}|\theta^*_a)\right]. \qquad [14]$$

For most applications, primary interest will center on the marginal posterior for θ^*_a. That is to say, inferences about latent number correct will typically be more important than direct statements about the individual π_{ia}.

The marginal distribution for θ^*_a may be obtained by integrating the joint posterior Equation 14 with respect to $\underset{\sim}{\pi}_a$, separately for each value of θ^*_a. Due to the first- and second-order independence assumptions of Equations 10 and 12, the desired multiple integral may be expressed as a product of single integrals. These will now be examined in more detail.

If $u_{ia} = 1$, the corresponding integral for π_{ia} will have the form

$$\int\pi_{ia}\, p(\pi_{ia}|\theta^*_a)\, d\pi_{ia} = E(\pi_{ia}|\theta^*_a)\ , \qquad [15]$$

the prior conditional expectation of π_{ia} given θ^*_a, which may be denoted by

$$\bar{P}_i(\theta^*_a) = E(\pi_{ia}|\theta^*_a)\ . \qquad [16]$$

Similarly, if $u_{ia} = 0$, the integral becomes

$$\int(1-\pi_{ia})p(\pi_{ia}|\theta^*_a)d\pi_{ia} = 1-\bar{P}_i(\theta^*_a) \qquad [17]$$

in the new notation. Combining the results of Equations 15 and 17 gives the following expression for the marginal posterior distribution for θ^*_a:

$$p(\theta^*_a | \underset{\sim}{u}_a) \propto p(\theta^*_a) \prod_i [\bar{P}_i(\theta^*_a)]^{u_{ia}} [1-\bar{P}_i(\theta^*_a)]^{1-u_{ia}} . \qquad [18]$$

Equation 18 may be seen as the most important single result in the present development and, as such, should be discussed in some detail. The most remarkable aspect of Equation 18, especially considering the minimal assumptions needed to obtain it, is its simplicity. Formally, it is identical to the result which is obtained beginning with the assumption that the π_{ia} are *known functions* of a postulated latent trait, using conditional independence of the u_{ia}, and introducing a prior distribution for the latent trait (see, for example, Birnbaum, 1969). In fact, in the present development, the latent trait (θ^*_a) is *constructed* in terms of the π_{ia}, making use only of the relative consistency assumption with regard to the latter.

Turning to the problem of specifying the prior density of Equation 11 for θ^*_a and $\underset{\sim}{\pi}_a$, Equation 18 makes it clear that only the marginal prior for θ^*_a and the conditional means $\bar{P}_i(\theta^*_a)$ need be considered. The precise form of the $p(\pi_{ia} | \theta^*_a)$ found in Equation 12 is irrelevant to the goal of making inferences about θ^*_a.

On the other hand, in specifying the conditional means, it should be recognized that the choices have implications for the corresponding conditional variances. In particular, given the restricted ranges of the π_{ia} as indicated in Equation 13, plus the inherent restriction to the interval (0,1), specifications of $\bar{P}_i(\theta^*_a)$ close to 0, π_m, or 1 imply that $\mathrm{Var}(\pi_{ia} | \theta^*_a)$ must be very small. If uncertainty regarding the true value of π_{ia} for a given value of θ^*_a is present in any substantial degree, it then follows that such "boundary" specifications should be avoided.

One concrete way of making appropriate specifications for $\bar{P}_i(\theta^*_a)$ will now be discussed. Suppose that a density from the beta family is chosen for $p(\pi_{ia} | \theta^*_a)$ and then truncated above or below π_m to conform with the restrictions in Equation 13. This choice has the advantages of providing a (conditional) prior for $\underset{\sim}{\pi}_a$, which is a natural conjugate to the likelihood of Equation 10, and (except for the truncation) of being familiar to users of elementary Bayesian techniques.

As described, for instance, in Novick and Jackson (1974), the choice may be completely specified by identifying the sum of the beta parameters ($\alpha + \beta$) with a hypothetical prior sample size (larger values reflecting more precise prior knowledge) and the mode of the density [$(\alpha - 1)/(\alpha + \beta - 2)$ for $\alpha, \beta > 1$] with the a priori most likely value for π_{ia}, given θ^*_a. The conditional means $\bar{P}(\theta^*_a)$ of the truncated densities may then be computed in terms of the *untruncated* mean ($\alpha/(\alpha + \beta)$) and incomplete beta functions as follows:

$$\bar{P}_i(\theta_a^*) = \begin{cases} \left(\frac{\alpha}{\alpha+\beta}\right) [1-I_{\pi_m}(\alpha+1,\beta)]/[1-I_{\pi_m}(\alpha,\beta)] & \text{for } i \leq \theta_a^*, \text{ and} \\ \left(\frac{\alpha}{\alpha+\beta}\right) I_{\pi_m}(\alpha+1,\beta)/I_{\pi_m}(\alpha,\beta) & \text{for } i > \theta_a^*. \end{cases} \qquad [19]$$

Here α and β are the parameters of the chosen beta density for π_{ia} given θ_a^* (and, thus, should formally have subscripts *i*, *a*, and θ_a^*).

The procedure just described provides a formal mechanism for "adjusting" prior estimates for π_{ia} away from the boundaries 0, π_m, and 1 to a degree which is controlled by the choice of hypothetical prior sample size ($\alpha + \beta$), with more precise prior knowledge leading to less adjustment. As mentioned earlier, the precise form of the truncated beta has no further consequences for inferences about θ_a^* and it should be recognized that there will be a range of densities having the same (or almost the same) modes and means as those obtained from any given choice of truncated betas.

Finally, a word should be said about the specification of $p(\theta_a^*)$. This will be a discrete distribution and it would seem wisest *not* to restrict its form in practice to that of any particular family, such as the binomial. It will often be the case that some information about the population distribution of ability in P is available from analyses carried out during the development of the test. Unless extra information is available regarding the specific person *a*, this population information will provide a good basis for specifying $p(\theta_a^*)$. In such a case, it may be useful to remember that Prob$(\theta_a^* \geq j)$ should ideally correspond to the proportion of persons in P who can master item *j* at level π_m.

Example of an Analysis

As noted at the beginning of this paper, the present treatment of abilities is most directly compatible with the work of Mokken (1971). In this spirit, Mokken and Lewis (1982) discuss an example in which the procedures proposed by the first author are applied for test development and evaluation, and those by the second author for the test administration phase (to make inferences about individual abilities). Suppose, however, that information regarding an existing test is only available in terms of a *parametric* item response model. What are the possibilities for analyzing individual responses to items in such a test using the approach described in the previous section? In what follows, this question will be considered in terms of a simple example.

Lord (1968) applied the 3-parameter logistic item response model,

$$\pi_{ia} = P_i(\theta_a) = \{1+\exp[-1.7a_i(\theta_a-b_i)]\}^{-1}(1-c_i) + c_i \,, \qquad [20]$$

to analyze the responses of nearly 3,000 persons to items of the Verbal Scholastic Aptitude Test. Most importantly for present purposes, he provided a table of estimated item parameters (a_i, b_i, and c_i) and of selected percentiles of the distribution of estimated θ_a values. While questions may be legitimately raised about various aspects of this analysis, as Lord himself did (for example, does the model provide a reasonable fit to the data?), the results are undoubtedly an important source of prior information for anyone wishing to work further with these items. In the present context, the problem becomes one of translating this information into the components required in Equation 18 to carry out an analysis of individual responses.

Remembering that θ_a^* may be thought of as dividing the θ_a scale into intervals, it would be convenient if the percentiles for θ_a given by Lord (1968) could be used to define boundaries for these intervals. In fact, the boundaries must be determined by individual items, relative to a given mastery level. For this example, an item mastery level $\pi_m = .6$ was chosen. This may be thought of as a correction of the "natural" value of .5 for "guessing," given that these are five-alternative multiple-choice items. Using Lord's tabled item parameters, items were then sought whose estimated response functions attained .6 for values of θ_a close to those given for reported percentiles. Since Lord provided parameters for 80 items of widely varying difficulty, this search ended successfully for six of the seven percentiles given (the exception being the .0025th). The results are summarized in Table 5.

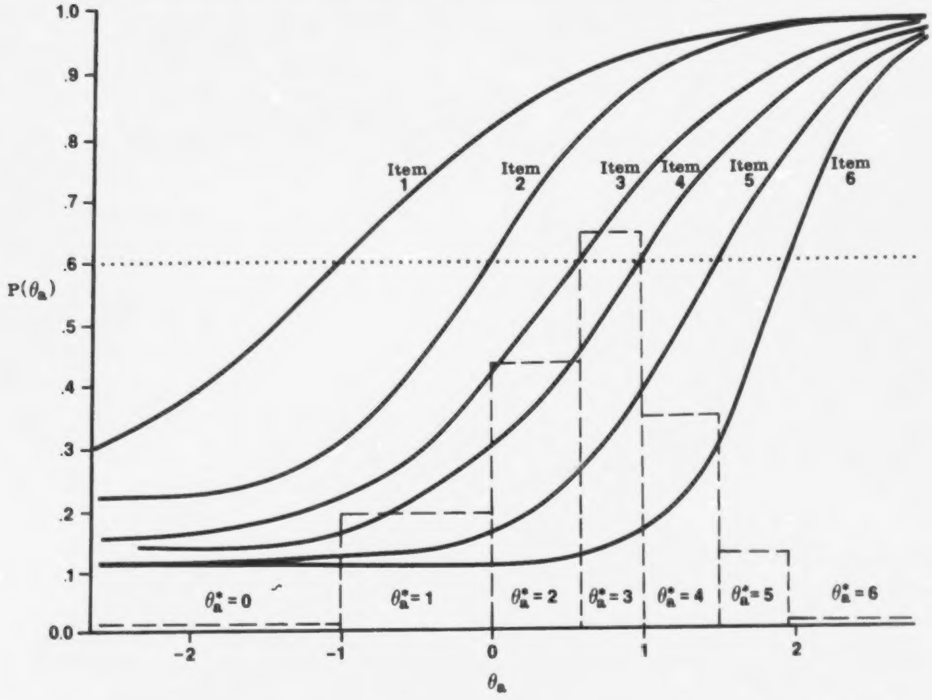
Table 5
Sample Items Chosen from Lord (1968)

Item No.	SAT No.	Reported Item Parameters a_i	b_i	c_i	θ_a Giving $P_i(\theta_a)=.6$	Reported Percentile
1	52	0.7	-1.0	.20	-1.00	.02
2	62	1.1	0.0	.20	0.00	.21
3	56	0.9	0.5	.14	0.59	.50
4	79	1.0	0.9	.13	0.99	.76
5	15	1.2	1.4	.11	1.50	.935
6	66	1.9	1.9	.11	1.96	.99

In the analysis which follows, only the six items described in Table 5 will be considered. This greatly simplifies presentation of results and makes it possible, as already indicated, to use the reported percentiles in selecting a prior distribution for θ_a^*. The estimated response functions for these items, obtained by substituting the parameter values from Table 5 in Equation 20, are shown in Figure 1, as is the relative frequency distribution for the estimated θ_a values, displayed as a histogram and based on the percentiles from Table 5.

Turning now specifically to the choice of priors, Table 6 gives the prior distribution for θ_a^* using the reported percentiles. Such a specification implies that the abilities of persons whose responses will be analyzed may be thought of as exchangeable with those for the group analyzed by Lord and that the abilities estimated in Equation 20 are thought to have a relationship to the

Figure 1
Item Response Functions and Distribution of θ_a from Table 5



θ^*_a defined by the π_{ia} for these six items and item mastery level of .6. Both of these implications refer to subjective evaluations rather than to objective facts regarding the π_{ia}. The requirement, thus, is that the implications be "reasonable," not that they be "true." In fact, as Lord (1968, p. 1018) has explained, for technical reasons, persons with extreme abilities (and especially low abilities) were over-represented in his sample. Thus, the prior of Table 6 is almost certainly too "low" and too "broad" for use with "typical" SAT examinees. This will be ignored in what follows.

Table 6
Prior Distribution for Latent Number Correct (θ^*_a),
Based on Data in Table 5, with $p(\theta_a) \times 100$

	Items						
Variable	1	2	3	4	5	6	
θ^*_a	0	1	2	3	4	5	6
Prior	2	19	29	26	17.5	5.5	1

The prior distribution for θ^*_a having been chosen, attention may be given to selecting conditional priors for the π_{ia} given θ^*_a. Following the approach of the previous section, truncated beta distributions will be considered for this purpose.

Most likely values (prior conditional modes) for the π_{ia} may be adopted from the estimated item response functions pictured in Figure 1. For simplicity, each function was evaluated at the midpoint of each of the finite θ_a intervals, corresponding to θ_a^* values 1 through 5. For the infinite intervals corresponding to θ_a^* of 0 and 6, the functions were evaluated at points with a distance equal to the average of the finite interval widths below the lowest and above the highest finite boundary, respectively.

The adoption of the actual sample size on which the estimates were based (2,862) as the hypothetical prior sample size for the betas might be considered. This would lead to conditional prior means for the π_{ia} virtually identical to the most likely values and would reflect complete confidence in the estimated functions obtained by Lord as descriptions of the π_{ia}. Such confidence is almost certainly not justified, and it may be argued that modal values close to 0, .6, and 1 required considerable adjustment to reflect uncertainty regarding the π_{ia} for an individual with ability θ_a^*. After some experimentation, a hypothetical prior sample size of 10 was adopted for all the conditional priors, and the required $\overline{P}_i(\theta_a^*)$ computed.

Figure 2 illustrates the process for Item 3 and $\theta_a^* = 2$. The truncated beta density for π_{3a} has its mode at the value given by the item response function $P_3(\theta_a)$ for θ_a at the midpoint of the interval corresponding to $\theta_a^* = 2$. Because of the truncation above .6, the conditional mean is somewhat lower than the mode, as indicated in the figure. The conditional means $\overline{P}_3(\theta_a^*)$ for the remaining values of θ_a^* are displayed as a step function in Figure 2 and indicate the extent of adjustment away from the boundaries obtained with a prior sample size of 10.

The full results are given in Table 7, where a "staircase" has been added to separate means for the items which, by definition, are mastered at each ability level from the rest. Thus, values above the staircase are all greater than the mastery level of .6 and those below all less than .6. A few extreme examples of adjustment from mode to mean are: .99 → .90 for Items 1 and 2, Ability 6; .11 → .19 for Items 5 and 6, Ability 0; .66 → .73 for Item 3, Ability 3; and .53 → .44 for Item 4, Ability 3.

Tables 6 and 7 provide all the prior information necessary to carry out via Equation 18 posterior analyses for individual abilities, based on responses to the six items of Table 5. Of course, it may reasonably be asked how much the prior distribution of Table 6 could possibly be modified by the information contained in only six responses. To shed some light on this question, consider first the posterior distributions based on two (hypothetical) extreme response records: all correct and none correct. These are shown in Table 8.

Figure 2
Adjustment from $P_3(\theta_a)$ to $\overline{P}_3(\theta^*_a)$, with $p(\pi_{3a}|\theta^*_a = 2)$ Shown Explicitly

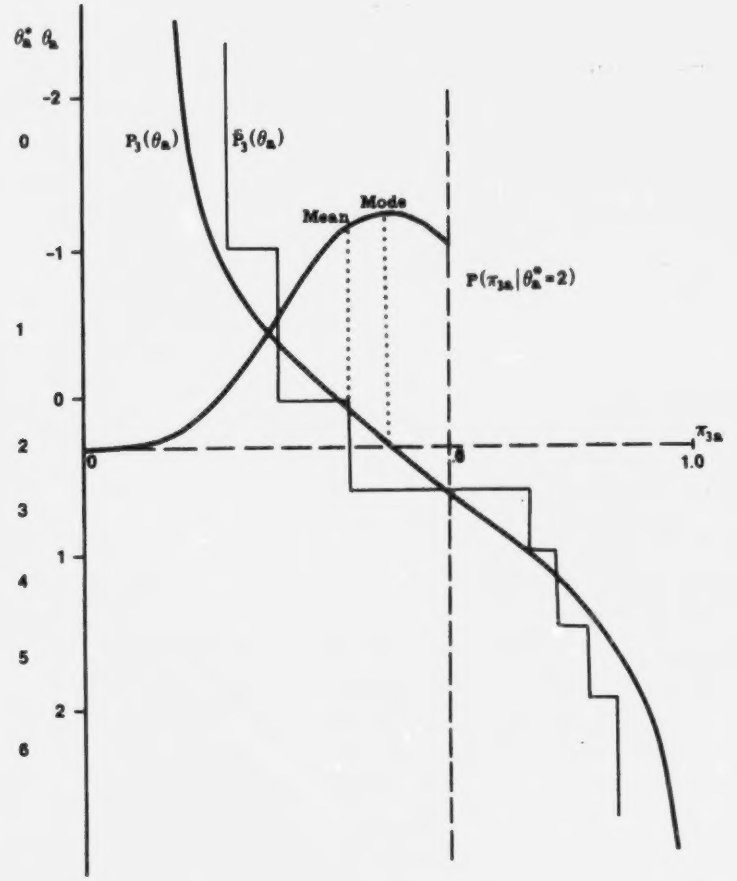


Table 7
Conditional Prior Means for π_{ia}
Given θ^*_a for Hypothetical Example,
with $\overline{P}_i(\theta^*_a) \times 100$

Items	θ* 0	1	2	3	4	5	6
1	42	75	81	85	87	88	90
2	28	40	74	81	85	88	90
3	24	32	43	73	78	83	88
4	21	26	36	44	74	80	87
5	19	20	25	33	43	74	85
6	19	19	19	21	26	40	84

The prior distribution for ability was chosen so that the probability assigned to $\theta_a^* \geq 3$ was just .5. It will be remembered that this may be interpreted as the a priori probability that person $\underline{a}$ can master Item 3 at the .6 level. How much has this been modified in the two distributions given in Table 8? The a posteriori probability that a person answering all six items correctly can master Item 3 is .945, the sum of the probabilities for $\theta_a^* = 3, 4, 5$, and 6. For a person answering none of the six items correctly, the probability that Item 3 can be mastered is only .04.

Table 8
Posterior Distributions for θ_a^* Based on Two Hypothetical Extreme Response Records ($\underset{\sim}{u}_a$), with $p(\theta_a^*|\underset{\sim}{u}_a) \times 100$

Record and Variable		Items											
		1		2		3		4		5		6	
θ_a^*	0		1		2		3		4		5		6
Record 1													
Responses		1		1		1		1		1		1	
Posterior	0		0.5		5		15		31		32		16.5
Record 2													
Responses		0		0		0		0		0		0	
Posterior	20		57.5		18.5		3.5		0.5		0		0

Thus, at least for these two extreme cases, it must be concluded that the responses have a substantial effect on the prior. That the prior also plays an important role in this example may be seen by noting that the posterior probability that a person answering all six items correctly can actually master all these items ($\theta_a^* = 6$, all $\pi_{ia} \geq .6$) is only .165. Similarly, the probability is only .20 that a person with none correct can actually master none ($\theta_a^* = 0$, all $\pi_{ia} < .6$).

As a final illustration of the interaction between prior and data in this example, sensitivity to response pattern will be considered. Once again, two hypothetical response records are analyzed, and the results are shown in Table 9. Each record consists of three correct and three incorrect responses. For simplicity, the first will be referred to as a "perfect" pattern and the second as having a single "reversal."

The perfect record suggests that the first three items can be mastered and, indeed, the highest posterior probability based on this record is assigned to $\theta_a^* = 3$. That the second highest probability is for an ability of 2, rather than 4, is a reflection of the prior, whose mode was 2 (see Table 6). What does the single reversal record suggest? Abilities of 2 and 4 seem more plausible than 3, since the latter implies two reversals rather than one. Comparing the resultant posterior distribution with the previous one, it may be observed that the probabilities for $\theta_a^* = 2$ and 4 are larger, while that for $\theta_a^* = 3$ is smaller.

Table 9

Posterior Distributions for θ^*_a Based on

Two Hypothetical Response Patterns, with $p(\theta_a | \underset{\sim}{u}_a) \times 100$

Record and Variable	Items												
		1		2		3		4		5		6	
θ^*_a	0		1		2		3		4		5		6
Record 1													
Responses		1		1		1		0		0		0	
Posterior	0.5		10		33		43		12.5		1		0
Record 2													
Responses		1		1		0		1		0		0	
Posterior	0.5		13		43.5		23		18		2		0

That the posterior distribution for ability based on a single reversal is unimodal rather than bimodal, and that the mode is 2 rather than 4, may be attributed to the influence of the prior.

Turning to the characteristic discussed for the extreme response records, namely, the posterior probability that Item 3 can be mastered, the perfect record results in a value of .565, while the single reversal leads to a probability of only .43. Thus, it may be concluded that within the present approach even small variations in response pattern can have a noticeable influence on inferences about abilities.

REFERENCES

Birnbaum, A. Statistical theory for logistic mental test models with a prior distribution of ability. Journal of Mathematical Psychology, 1969, 6, 258-276.

Cliff, N. Test theory without true scores? Psychometrika, 1979, 44, 373-393.

Lewis, C. Estimating abilities: Inferences for random variables. Kwantitatieve Methoden, 1981, 2, 17-34.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lumsden, J. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 1978, 31, 19-26.

Mokken, R. J. A theory and procedure of scale analysis with applications in political research. New York: de Gruyter/Berlin: Mouton, 1971.

Mokken, R. J. Personal communication, 1982.

Mokken, R. J., & Lewis, C. A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 1982, 6, 417-430.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Stokman, F. N., Van Schuur, W. H. Basic scaling. Quality and Quantity, 1980, 14, 5-30.

Tatsuoka, K. K., & Tatsuoka, M. M. Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Psychology, in press.

Technisch Centrum FSW. STAP user's manual (Vol. 4. Stochastic cumulative scaling: Mokken scale, Mokken test). Amsterdam: University of Amsterdam, 1980.

DISCUSSION

ROBERT K. TSUTAKAWA
UNIVERSITY OF MISSOURI

There was a time when statisticians regarded Bayesian and nonparametric statistics as incongruent. I am happy to see that Lewis has brought these two areas together very forcefully. One general principle applies to Lewis's paper: Sensible modeling with the proper use of subjective information will generally result in good procedures when the sample size is small to moderate. Here, discussion will be limited to (1) some amplification of the method proposed, (2) difficulties with the method, and (3) suggestions for further work.

The basic components of this paper consist of a finite set of $\underline{n}$ test items, a finite set of $\underline{k}$ examinees, and a response matrix $\underline{u} = \{u_{ia}\}$ consisting of binary responses where $u_{ia} = 0$ or 1 according to whether examinee $\underline{a}$ responds to item $\underline{i}$ incorrectly or correctly. The true score of person $\underline{a}$ to item $\underline{i}$ is defined by $\pi_{ia} = E(u_{ia})$, or the probability that $u_{ia} = 1$. For a given mastery level π_m, $0 < \pi_m < 1$, the latent number-correct score θ_a^* for person $\underline{a}$ is defined as the number of items for which $\pi_{ia} \geq \pi_m$.

Lewis has introduced a new type of ordering of the items. The ordering is such that given two items, $\underline{i}$ and $\underline{j}$, $i < j$ if and only if for any person $\underline{a}$, if $\pi_{ja} \geq \pi_m$, then $\pi_{ia} \geq \pi_m$ for every $i < j$. This ordering and score θ_a^* is illustrated in Figure 1.

Note that this ordering of the true scores in Figure 1 is not a linear ordering of $\{\pi_{1a}, \ldots, \pi_{na}\}$; for example, if $\theta_a^* = j$, then π_{ja} need not be the smallest among $\{\pi_{ja}; \pi_{j-1,\, a}, \ldots, \pi_{1a}\}$. However, if $\theta_a^* = j - 1$, then π_{ja} is the smallest among $\{\pi_{ja}, \pi_{j-1,a}, \ldots, \pi_{1a}\}$. Thus, the restriction among $\{\pi_{1a}, \ldots, \pi_{na}\}$ is dependent on θ_a^*.

In order to visualize this new ordering, it is instructive to picture it in terms of item response curves. Figure 2 shows three curves and the value of θ_a^* for the different intervals of ability. The items have been labeled 1, 2, and 3 according to the new ordering. This figure demonstrates the following important property of the ordering:

> If the item response curves are monotonically increasing with respect to ability, then the value of θ_a^* is nondecreasing with respect to ability.

Figure 1
Item Ordering Implied by Lewis's Procedure

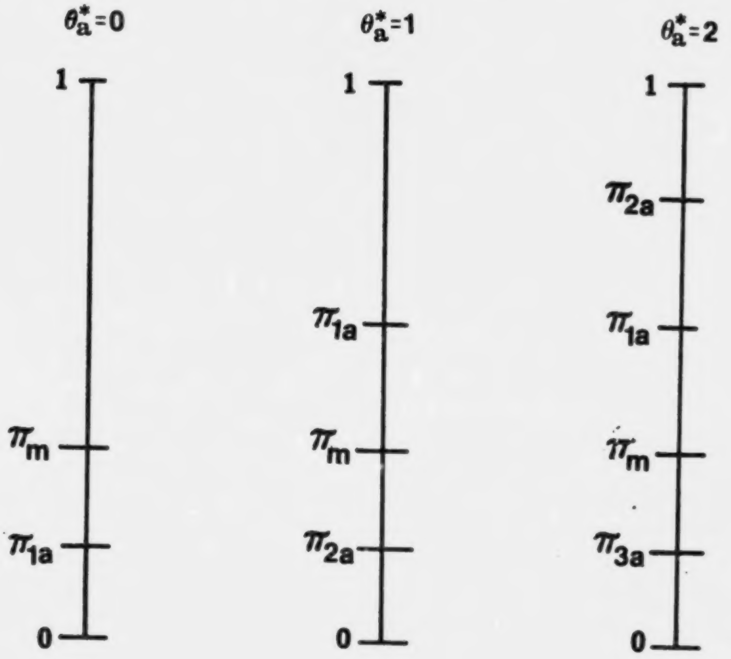
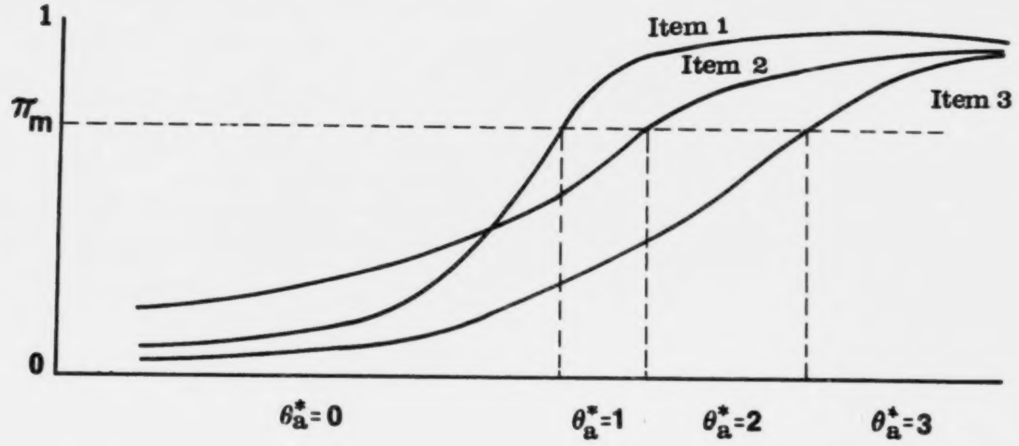

θ*a=0
1
πm
π1a
0
θ*a=1
1
π1a
πm
π2a
0
θ*a=2
1
π2a
π1a
πm
π3a
0

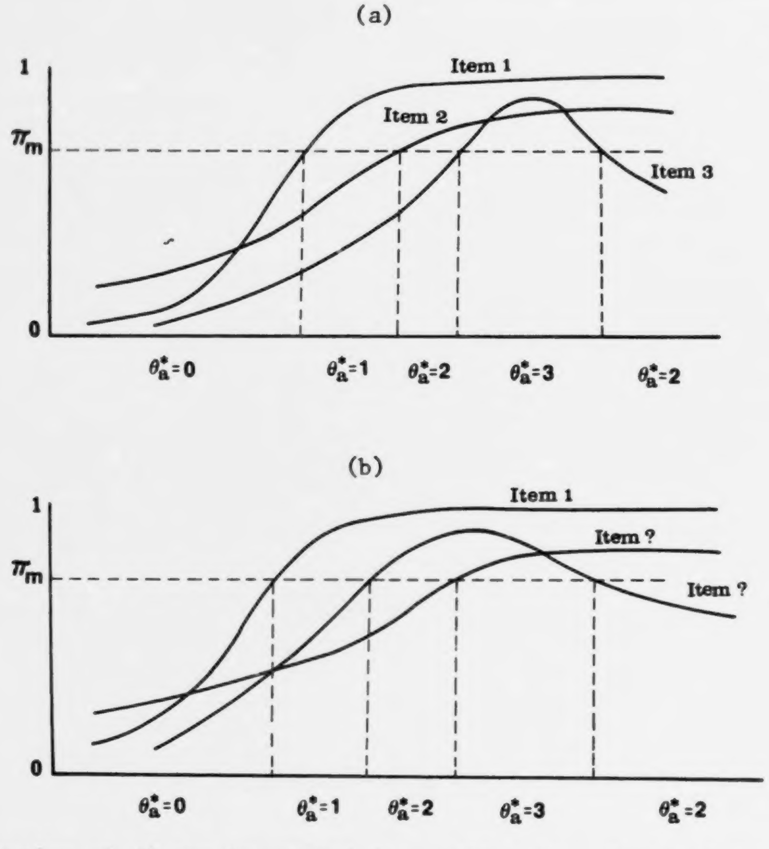
Figure 2
Item Response Curves When Monotonicity Holds



When item response curves are monotonically increasing, the ordering is in terms of the order in which it crosses the mastery level π_m. What happens before and after the crossing is irrelevant in this case. It is convenient to think of θ_a^* as a discretized ordering of ability, which is usually measured on a continuous scale.

To see what can happen when the monotonicity condition is violated, consider Figures 3a and 3b. It can be seen that the items can still be ordered in Figure 2 but not in Figure 3. These figures also show that θ_a^* is not nondecreasing with respect to ability.

Figure 3
Item Response Curves with Violations of Monotonicity



The estimation of θ_a^* can be viewed from either the classical or Bayesian approach. Let $\underset{\sim}{u}_a' = (u_{1a}, \cdots, u_{na})$ denote the response to the $\underline{n}$ items by person $\underline{a}$. The classical approach is deductive and considers the probability of $\underset{\sim}{u}_a$ for different latent scores $\pi_a' = (\pi_{1a}, \cdots, \pi_{na})$ and from this constructs an estimate of θ_a^*. The Bayesian approach which Lewis adopts begins with a prior distribution of θ_a^* and derives the posterior distribution of θ_a^* given $\underset{\sim}{u}_a$. Under the usual assumption of local independence, Lewis shows that since the posterior is

$$P(\theta_a^* | \underset{\sim}{u}_a) \propto P(\theta_a^*) \prod_{i=1}^{n} \bar{P}_i(\theta_a^*)^{u_{ia}} [1 - \bar{P}_i(\theta_a^*)]^{1-u_{ia}} , \qquad [1]$$

where

$$\bar{P}_i(\theta_a^*) = E(\pi_{ia} | \theta_a^*) , \qquad [2]$$

the only specification needed from the prior is the prior expectation of π_{ia} for different values of θ_a^*. Thus, once the ordering assumption is made and the prior expectations are available, the computation of the posterior of θ_a^* becomes routine.

In applying this type of approach to specific items and persons, it is important to know how much of the inference is subjective and how to express a prior opinion about the latent true score. In the context of item response curves which I have discussed, it would seem relatively easy to assume the existence of an ordering in terms of a mastery level. It is considerably more difficult to specify the prior $p(\underset{\sim}{\pi}_a^*)$ or even $p(\theta_a^*)$ without a fair amount of knowledge about each item and person. Working with the truncated beta on $\underset{\sim}{\pi}_a$, as shown by Lewis, is more a mathematical convenience than a method which can be used with conviction. The difficulty seems to be the manner in which the conditional distribution of $\underset{\sim}{\pi}_a$, given θ_a^*, varies with θ_a^* and the choice of the prior parameters needed to specify the truncated beta. In particular, there is no simple interpretation of these parameters as there is in the case of the non-truncated beta.

Although the concept of latent number-correct score does not depend on the idea of ability commonly used in mental testing, it would be interesting to relate $\underset{\sim}{\pi}_a$ and θ_a^* to item response curves. Since these curves satisfy the ordering assumption (under mild conditions), a prior distribution on these curves together with a prior distribution on ability will define a prior on $\underset{\sim}{\pi}_a$ and therefore on θ_a^*.

From a Bayesian perspective, however, it is not necessary to go beyond the specific set of items and persons. The selection of the prior distribution will be facilitated by considering a larger universe, possibly a hypothetical one, from which the given persons are randomly selected. This would suggest a population interpretation of the distribution of θ_a^* (for a set of persons) and of the underlying $\underset{\sim}{\pi}_a$. Looking at this larger context may provide a formal mechanism for using previous experiments to get a better labeling of the items and prior distribution of $\underset{\sim}{\pi}_a$ in specific cases.

UNIDIMENSIONAL AND MULTIDIMENSIONAL MODELS FOR ITEM RESPONSE THEORY

RODERICK P. MCDONALD
MACQUARIE UNIVERSITY

McDonald (1967a, 1967b) described a very general model for nonlinear common factor analysis and showed that its special cases include the major models in latent trait, or item response, theory. Also, Lord and Novick's (1968) widely influential account of item response theory (IRT) sets out a number of relationships betweeen latent trait theory and (linear or nonlinear) common factor theory. Nevertheless, it seems that the substantial mathematical overlap between IRT and common factor theory is still not widely appreciated. The object of this paper is to provide an up-to-date review of the relationship between IRT and (nonlinear) common factor theory and to draw out of this relationship some implications for current and future research in IRT.

At the outset, there seems to be a notable difference between latent trait theory and common factor theory. Latent trait theory is defined by the principle of local independence, while common factor theory is based upon a weaker principle. In the context adopted in this paper, that of test theory, the principle of local independence states that conditional upon any fixed value of a (scalar or $\underline{k}$-component vector) latent trait $\underset{\sim}{\theta}$ associated with the members of a population of examinees, the scores of the examinees on a set of $\underline{n}$ items or tests are mutually statistically independent. The principle actually constitutes the definition of the latent trait(s) $\underset{\sim}{\theta}$. That is, $\underset{\sim}{\theta}$ is a vector of latent traits if and only if it is a set of quantities associated with the examinees such that the responses are mutually statistically independent for any fixed value of $\underset{\sim}{\theta}$. The English language labels "latent," "trait," and commonly associated words such as "underlying," "unobservable," and "ability" do not contain or convey any part of this definition and do not add anything to it. Since the latent traits explain all associations between the item responses, a latent trait is ordinarily interpreted as a measure of a trait in the psychological sense of the word, that is, an abstract attribute of the examinees which the items measure in common. In particular, in the case of a cognitive test the trait is commonly interpreted as an "ability."

Common factor theory, together with its common factor scores, is defined by the principle that conditional upon any value of a $\underline{k}$-component vector $\underset{\sim}{\theta}$ of common factor scores, the scores of examinees on a set of items or tests are mutually uncorrelated. McDonald (1979, 1981) has suggested that the principle of local independence is described as taking two forms; in the strong form the responses are conditionally independent, and in the weak form they are conditionally uncorrelated. It is reasonable to assert that in all work with latent traits or common factors the strong form of the principle is assumed, although ordinarily the implied weak form, at most, is tested verifying that the residual covariances of the items are acceptably small but not checking whether the higher moments of their joint distribution are, conditionally, products of their univariate moments. Unfortunately, in a great deal of work on IRT (and occasionally in common factor analysis) the investigators do not even obtain the residual covariances of the items, or any other statistic such as a function of the residual covariances which measures the extent to which the defining principle of the model is satisfied. Since it is always easy to compute the residual covariances, it is difficult to understand why this is not standard practice. Possibly a failure to examine residual covariances in IRT models follows from a failure to recognize them as common factor models. If so, it can immediately be suggested that a central implication of the factor analytic perspective on IRT is that it should always prove informative to compute and examine the residual covariance matrix of the items when fitting such "standard" models as the Rasch model, the normal ogive model, and the logistic model, as well as other models to be discussed below.

The general nonlinear common factor model (McDonald, 1962a, 1967a, 1967b) may be expressed as the nonlinear multivariate regression equation

$$\hat{\underset{\sim}{y}} = \mathcal{E}\{\underset{\sim}{y}|\underset{\sim}{\theta}\} = \underset{\sim}{\phi}(\underset{\sim}{\theta}) \qquad [1]$$

$$\hat{y}_j = \mathcal{E}\{y_j|\theta\} = \phi_j(\theta),\ j = 1, \ldots, n, \qquad [2]$$

where

$\underset{\sim}{y}' = [y_1, \cdots, y_n]$ is a vector of test or item scores;

$\underset{\sim}{\phi}$ is a vector of functions, in general, nonlinear; and

$\underset{\sim}{\theta}$ is a k × 1 vector of latent traits or common factors defined by the principle of local independence, at least in its weak form.

A vector of residuals $\underset{\sim}{e}$ is defined by

$$\underset{\sim}{e} = \underset{\sim}{y} - \mathcal{E}\{\underset{\sim}{y}|\underset{\sim}{\theta}\} \qquad [3]$$

or

$$e_j = y_j - \mathcal{E}\{y_j|\underset{\sim}{\theta}\},\ j = 1, \ldots, n\ . \qquad [4]$$

It then follows without further assumptions that the covariance structure of the observed scores is given by

$$\underset{\sim}{C} = \mathrm{Cov}\{\underset{\sim}{y}\} = \mathrm{Cov}\ \{\underset{\sim}{\phi}\} + \underset{\sim}{U}^2\ , \qquad [5]$$

where $\underset{\sim}{U}^2$, diagonal, nonnegative definite, is the residual covariance matrix and contains unique variances in the usual sense of the term in common factor theory (see McDonald, 1967b). The residual is ordinarily composed of a stable specific part and an unstable error of replication. It is important to recognize that, in general, the residual variances are not required to be homoscedastic and that binary items, as well as "quantitative" test scores, have both a specific and an error component in the sense understood from common factor theory.

McDonald (1982b, in press) suggests a fundamental classification of the family of all latent trait or common factor models into (1) *strictly linear models*, in which the functions $\phi(\boldsymbol{\theta})$ are linear both in the coefficients of the regressions on the latent traits and in the latent traits themselves; (2) *wide-sense linear models*, in which the functions $\boldsymbol{\phi}(\boldsymbol{\theta})$ are linear in the coefficients of the regressions but not in the latent traits; and (3) *strictly nonlinear models*, which, loosely speaking, are nonlinear in both. It is safer to define a strictly nonlinear model as one that cannot be expressed as a wide-sense linear model with a finite number of terms, since any strictly nonlinear model may in principle be approximated as closely as desired by a wide-sense linear model using harmonic analysis (McDonald, 1967a).

The linear common factor model

$$\hat{y}_j = \sum_{p=1}^{k} f_{jp}\theta_p \qquad [6]$$

is the obvious example of a strictly linear model. Lazarsfeld's (Lazarsfeld & Henry, 1968) latent class model is also strictly linear.

The exploratory methods for nonlinear factor analysis described by McDonald (1962b, 1967b, 1967c, 1967d) essentially consist of wide-sense linear models, in which the regression function is a linear combination of polynomial functions, and, possibly, products of latent traits. A central result of the theory is the demonstration (McDonald, 1967b) that if a wide-sense linear model contains r functions of the k latent traits, then the covariance structure of the observed variables can be expressed as

$$\mathbf{C} = \mathrm{Cov}\{\mathbf{y}\} = \mathbf{B}\mathbf{B}' + \mathbf{U}^2 \qquad [7]$$

where $\mathbf{B}$ is an $n \times r$ matrix. Consequently, the model cannot be distinguished from the strictly linear model with r common factors on the basis of the covariance structure alone. This fact serves to explain the problem of difficulty factors in the factor analysis of binary items (McDonald, 1967a; McDonald & Ahlawat, 1974).

The best-known examples of strictly nonlinear latent trait models would be the almost interchangeable normal ogive and logistic models of IRT. Other less well-known examples are the *latent distance models* and *latent content models* (Lazarsfeld & Henry, 1968). Neither of these, however, seems of interest in the context of test theory.

In principle, any latent trait model can be treated as a *random-regressors model* or a *fixed-regressors model*, the choice being dictated by a number of mathematical considerations. The least of these, curiously enough, is the fundamental logic of the distinction. In the random regressors model, the n observed variables $y_1, \ldots, y_n$ and the k latent traits $\theta_1, \ldots, \theta_k$, are assumed to constitute $n + k$ random variables, with a joint distribution in $n + k$ dimensions. A sample of size N drawn from this distribution consists of N random vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$ independently and identically distributed like $\mathbf{y}$, and there is no question of estimating the unknown corresponding $\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_N$. In the fixed-regressors model, a sample of N observations $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_N$ of $\underset{\sim}{y}$ is thought of as corresponding to N fixed values, $\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_N$ of $\underset{\sim}{\theta}$, which are estimated jointly with the coefficients of the regression. From one point of view, there is an n × N matrix sample of size one from a peculiar conceptual universe that consists of all sets of N examinees with the same (unknown) values of $\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_N$ as in the sample that has been drawn.

In the terminology of Neyman and Scott (1948), the coefficients of the regression are *structural parameters*, while the latent trait values are *incidental parameters* whose number increases with N. Neyman and Scott showed that in a model containing such incidental parameters, the structural parameters may not be consistently estimated. In the linear case of the analogous multivariate regression model in which the independent variables are known quantities, the choice between a fixed-regressors model and a random-regressors model rests essentially on an application of the logic of the distinction. That is, the decision depends on whether the values of the independent variables are in fact randomly drawn with the experimental units or fixed by experimental control. (Some investigators seem unaware of the distinction and apply the statistics of the fixed-regressors model when they have random independent variables.) In the nonlinear case, the investigator will be likely to use the fixed-regressors model, even when the logic of the research design calls for random regressors, simply because the nonlinear random-regressors theory is difficult or impossible to work out. It is this latter consideration that seems to determine the choice in latent trait theory, where the regressor variables are defined by the principle of local independence. It seems unlikely, otherwise, that employing a fixed-regressors model in latent trait theory would ever be considered, since, rather obviously, a randomly drawn examinee brings his/her latent trait value(s) with him/her, so to speak.

Random-regressors and fixed-regressors treatments of latent trait theory each have their virtues and limitations. The next section examines some random-regressors models, while the section following it concerns fixed-regressors models.

Random-Regressors Theory

The first effective treatment of a latent trait model as a random-regressors model is possibly Lord's (1952) demonstration that if a set of binary items fits the normal ogive model and the latent trait is a normally distributed random variable, then the tetrachoric correlations of the items can be explained by the Spearman single common factor model. It is an open question whether the "heuristic" method, based on this result, of estimating the item parameters of the normal ogive model using a Spearman factor analysis of tetrachoric correlations has been or can be substantially improved upon, in terms of actual numerical results.

McDonald (1967a) pointed out that a (strictly) nonlinear latent trait model such as the normal ogive and latent distance models can be approximated as closely as desired by a polynomial series (a wide-sense linear model) on the basis of harmonic analysis (Fourier analysis), a traditional device in the physical sciences for reducing a nonlinear model to a linear model. The coefficients of the polynomial series are so chosen that any finite segment of it is a least squares approximation to the desired strictly nonlinear model, weighted by the density function of the latent trait.

The main application of this method was to the normal ogive model, under the assumption that the latent trait is a random regressor with a normal distribution. In this case, McDonald (1967a) showed that if the normal ogive model is written as

$$\hat{y}_j = N(f_{j0} + f_{j1}\theta) \qquad [8]$$

where $N(\cdot)$ is the (unit) cumulative normal distribution function, then a weighted least-squares approximation to the model is given for any choice of r by

$$\phi_j^{(r)} = \sum_{p=0}^{r} b_{jp} h_p(\theta) \qquad [9]$$

where $h_p(\theta)$ is the normalized Hermite-Tchebycheff polynomial of degree p, given by

$$h_p(\theta) = \frac{1}{\sqrt{p!}} \sum_{t=0}^{q} (-)^t \frac{\theta^{p-2t}}{2^t t! p-2t!} \qquad [10]$$

with $q = s$ if either $p = 2s$ or $p = 2s + 1$

$$b_{j0} = N\{f_{j0}/(1 + f_{j1}^2)^{1/2}\} \qquad [11]$$

and

$$b_{jp} = \frac{1}{\sqrt{p}} \left[\frac{f_{j1}}{(1 + f_{j1}^2)^{1/2}}\right]^p h_{p-1}\left[\frac{f_{j0}}{(1 + f_{j1}^2)^{1/2}}\right] n\{f_{j0}/(1 + f_{j1}^2)^{1/2}\} \qquad [12]$$

where $n(\cdot)$ is the (unit) normal density function. McDonald (1982a) has shown that the normal ogive is well approximated by the cubic polynomial obtained by terminating the series Equation 9 at $p = 3$. McDonald (1967a) proposed to fit the normal ogive model by estimating the first few coefficients of the series Equation 9 by nonlinear factor analysis, and solving sample analogues of equations 11 and 12 for f_{j0}, f_{j1}. Unpublished work by McDonald and Ahlawat (1974) showed that satisfactory estimates of f_{j0}, f_{j1} can be obtained from a (linear) Spearman factor analysis of the item covariance matrix to yield estimates of the quantities b_{j0}, b_{j1} (without obtaining coefficients of higher degree terms) and simultaneously solving the two equations 11 and

$$b_{j1} = \left\{\frac{f_{j1}}{(1 + f_{j1}^2)^{1/2}}\right\} n \left\{\frac{f_{j0}}{(1 + f_{j1}^2)^{1/2}}\right\} \qquad [13]$$

for f_{j0}, f_{j1} (substituting sample analogues). Thus, from this work a second "heuristic method" based on a Spearman factor analysis of the items is obtained.

Bock and Lieberman (1970) were able to obtain maximum likelihood estimates of the item parameters f_{j0}, f_{j1} in the normal ogive model, again assuming that the latent trait is a random regressor with a normal distribution. The method used a point approximation to the continuous distribution of the latent trait to yield a direct attack on the evaluation of the required integrals. The method was limited in application by the heavy computing demands in obtaining these approximations.

Christoffersson (1975) treated what may be recognized as a random-regressors multidimensional normal ogive model, using theory having elements in common with that of McDonald (1967a) and of Bock and Lieberman (1970). Christoffersson (1975) defined the model by introducing a set of n theoretical continuous variables that fit a multiple common factor model, and by supposing that the n observed binary item responses arise by dichotomization of these theoretical continuous variables at n threshold values to be estimated along with the parameters of the "underlying" common factor model. An expression is easily obtained for the proportion of examinees passing each item as a function of the model parameters. The proportion passing a pair of items can be approximated as closely as desired by taking sufficient terms of Pearson's tetrachoric series. (Christoffersson chose to stop at 10 terms.) Sample estimates of these proportions were then used to fit the model by generalized least squares. Muthén (1978) suggested the generalized least squares estimators in Christofferson's (1975) model can be obtained with less numerical cost by using sample tetrachoric correlations in place of sample proportions passing pairs of items.

McDonald (1980, 1982a, 1982b) has shown that the unidimensional normal ogive model can be fitted by ordinary least squares, as an application of a general purpose program for the analysis of covariance (moment) structures, using the harmonic analysis results of Equations 8 through 13. Instead of first fitting a Spearman model to the item covariances and then solving sample analogues of Equations 11 and 13 for the item parameter estimates, the covariance structure implied by Equation 9, with the coefficients given by Equations 11 and 12, is fit minimizing the least squares function with respect to the fundamental parameters f_{j0}, f_{j1}, $j = 1, \ldots, n$.

So far this section has consisted essentially of a review of known results. The remainder of this section will concern an extension of McDonald's (1967a) harmonic analysis of the unidimensional normal ogive to the multidimensional case and will relate this theory to Christoffersson's (1975) dichotomization of the common factor model. A multidimensional normal ogive model with a linear combination rule for the latent traits is defined as

$$\hat{y}_j = N\{f_{j0} + f_j'\theta\} = N\{f_{j0} + f_{j1}\theta_1 + \ldots + f_{jk}\theta_k\}, \qquad [14]$$

where, as before, $N(\cdot)$ is the cumulative normal distribution function. It is assumed that $\boldsymbol{\theta}$ is a random regressor with a k-variate normal distribution, and a metric is chosen such that each component of θ has mean zero and variance unity. $\mathbf{P}$ is written for the covariance (correlation) matrix of $\boldsymbol{\theta}$ and

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1' \\ \vdots \\ \mathbf{f}_n' \end{bmatrix} \qquad [15]$$

where $\mathbf{f}_j' = [f_{j1}, \ldots, f_{jk}]$. In some applications a pattern may be prescribed for F, with elements constrained to be zero (to define a "simple structure") or subject to other desired constraints, as in the counterpart factor-loading matrix in multiple factor analysis. In particular, by constraining the nonzero elements in each column to be equal (while choosing a pattern such that they are linearly independent), a multidimensional counterpart of the Rasch model is obtained. For each item score y_j, the item characteristic function $N(f_{j0} + \mathbf{f}_j'\boldsymbol{\theta})$ is constant on planes of dimension k - 1 orthogonal to the vector $\mathbf{f}_j$. In the Appendix it is shown in detail that Equation 14 may be represented by the infinite polynomial series.

$$\hat{y}_j = \phi_j^{(\infty)}(\boldsymbol{\theta}) = \sum_{p=0}^{\infty} b_{jp} h_p \left(\frac{1}{d_j} \mathbf{f}_j'\boldsymbol{\theta}\right) , \qquad [16]$$

where b_{j0}, b_{jp} are obtained by substituting

$$d_j = (\mathbf{f}_j'\mathbf{P}\mathbf{f}_j)^{1/2} \qquad [17]$$

for f_{j1} in Equations 11 and 12, and $h_p(')$ is given by Equation 10 as before. The first r terms of the series Equation 16, which will be denoted by $\phi_j^{(r)}(\boldsymbol{\theta})$, yield a polynomial of degree r, which, like the multidimensional normal ogive Equation 14, is constant on planes of dimension k - 1 orthogonal to the vector $\mathbf{f}_j$ and yields a weighted least squares approximation to it, as in the unidimensional case treated by McDonald (1967a). It further follows that the proportion of examinees passing an item is given by

$$\boldsymbol{\pi}_j = \mathbf{b}_{j0} \qquad [18]$$

and the proportion passing two items j and k is given by

$$\boldsymbol{\pi}_{jk}^{(\infty)} = \sum_{p=0}^{\infty} b_{jp} b_{kp} \left(\frac{1}{d_j} \mathbf{f}_j'\mathbf{P}\mathbf{f}_k \frac{1}{d_k}\right)^p , \qquad j \neq k = 1, \ldots, n . \qquad [19]$$

In practice the first r terms of Equation 19 may be substituted, denoted by $\pi_{jk}^{(r)}$, as a finite approximation to it. The model may be fit by least squares, minimizing the squared differences between sample proportions p_{jk} passing items j and k, and $\pi_{jk}^{(r)}$ given by Equation 19, with respect to f_{j0}, $\mathbf{f}_j$.

In the unidimensional case, Equation 19 reduces to the simple form

$$\pi_{jk}^{(\infty)} = \sum_{p=0}^{\infty} b_{jp}\, b_{kp}, \quad j \neq k = 1, \ldots, n \,, \tag{20}$$

where b_{jp} is given by Equations 11 and 12, and the moment structure Equation 20 is then easily fitted by an application of McDonald's (1978, 1980) comprehensive model for the analysis of covariance structures, using a facility in the computer program for this work to evaluate b_{jp} and corresponding derivatives, as functions of f_{j0}, f_{j1}. In contrast, the moment-structure Equation 19 does not seem to yield an advantageous matrix expression and is comprehended by the comprehensive model only in the sense that the bivariate moments themselves are programmable functions of the model parameters. The simple linear approximation

$$\pi_{jk}^{(1)} = b_{j0}\, b_{k0} + \left[\frac{b_{j1}}{d_j}\right] \mathbf{f}_j' \mathbf{P}\, \mathbf{f}_k \left[\frac{b_{k1}}{d_k}\right], \quad j \neq k = 1, \ldots, n \,, \tag{21}$$

may be written in matrix form as

$$[\pi_{jk}^{(1)}] = \mathbf{b}_0 \mathbf{b}_0' + \mathbf{D}^{-1}\, \mathbf{D}_b \mathbf{F} \mathbf{P} \mathbf{F}' \mathbf{D}_b \mathbf{D}^{-1} + \mathbf{D}_u \tag{22}$$

where

$\mathbf{b}_0' = [b_{10}, \ldots, b_{n0}]'$,

$\mathbf{D} = \text{Diag}\ \{d_1, \ldots, d_n\}$,

$\mathbf{D}_b = \text{Diag}\ \{b_{11}, \ldots, b_{n1}\}$, and

$\mathbf{D}_u = \text{Diag}\ \{b_{10}(1-b_{10}) - b_{11}^2, \ldots, b_{n0}(1-b_{no}) - b_{n1}^2\}$.

Monte carlo studies of the unidimensional case have shown that the precision of estimation of the item parameters does not depend upon the number of terms retained in the polynomial approximation. On the other hand, the residuals $p_{jk} - \pi_{jk}^{(r)}$ tend, of course, to reduce in absolute values as r is increased, though convergence is very rapid, and terms beyond the cubic in general seem negligible. A simple and probably numerically efficient way to fit the model is to fit the linear approximation Equation 21 in the form Equation 22 using McDonald's general model for the analysis of covariance structures and then to use Equation 19 to some suitable order in order to evaluate the residuals more precisely.

However, at this point it should be noted that the theory just described is much closer to that of Christoffersson (1975) than might be obvious at first sight. Indeed, the expression Equation 19 is identical with Equation 27 below, given by Christoffersson, except for a reparameterization of the model. Christoffersson defined a set of unobservable variables, $\underset{\sim}{v}$, that follow the multiple common factor model, say,

$$v_j = \underset{\sim}{\lambda}_j' \underset{\sim}{\theta} + \delta_j \qquad [23]$$

where

$$\underset{\sim}{\Lambda} = \begin{bmatrix} \underset{\sim}{\lambda}_1' \\ \vdots \\ \underset{\sim}{\lambda}_n' \end{bmatrix} \qquad [24]$$

is a n × k matrix of common factor loadings,
$\underset{\sim}{\theta}$ is a vector of common factors, and
δ_j is the *j*th unique factor.

He then supposed that

$$\begin{aligned} y_j &= 1 \quad \text{if } v_j \geq t_j \\ &= 0 \quad \text{if } v_j < t_j \, . \end{aligned} \qquad [25]$$

It then follows that the proportion of examinees passing item *j* is given by

$$\pi_j = N(t_j) \qquad [26]$$

and the joint proportion of examinees passing items *j* and *k* is given by

$$\pi_{jk} = \sum_{p=0}^{\infty} \tau_p(t_j)\tau_p(t_k)(\underset{\sim}{\lambda}_j' P \underset{\sim}{\lambda}_k)^p \, , \qquad [27]$$

where $\tau_p(\cdot)$ is the tetrachoric function defined by

$$\tau_0(t_j) = N(t_j) \qquad [28]$$

$$\tau_p(t_j) = \frac{1}{\sqrt{p}} h_{p-1}(t_j) n(t_j) \, , \quad p = 1, 2, \ldots, \qquad [29]$$

and, as before, $h_p(\cdot)$ is the normalized Hermite-Tchebycheff polynomial Equation 10 and $n(\cdot)$ is the normal density function. Using Equations 28 and 29, the result (Equation 19) obtained by harmonic analysis of the model Equation 14 is rewritten as

$$\pi_{jk}^{(\infty)} = \sum_{p=0}^{\infty} \tau_p \left[\frac{f_{j0}}{(1+d_j^2)^{1/2}}\right] \tau_p \left[\frac{f_{k0}}{(1+d_k^2)^{1/2}}\right] \left[(1+d_j^2)^{-1/2} \utilde{f}_j' \utilde{P} \utilde{f}_k (1+d_k^2)^{-1/2}\right] \qquad [30]$$

It is then immediately evident that the expressions Equations 27 and 30 are identical except for a choice of parameterization. Each is obtained from the other by writing

$$f_{j0} = \frac{t_j}{\sqrt{1-\utilde{\lambda}_j' \utilde{P} \utilde{\lambda}_j}} \; ; \; \utilde{f}_j = \frac{\utilde{\lambda}_j}{\sqrt{(1-\utilde{\lambda}_j' \utilde{P} \utilde{\lambda}_j}} \qquad [31]$$

or, conversely,

$$t_j = \frac{f_{j0}}{\sqrt{1+\utilde{f}_j' \utilde{P} \utilde{f}_j}} \; ; \; \utilde{\lambda}_j = \frac{\utilde{f}_j}{\sqrt{1+\utilde{f}_j' \utilde{P} \utilde{f}_j}} \; . \qquad [32]$$

To put this another way, Christoffersson's work implies writing the multidimensional normal ogive model as

$$\hat{y}_j = \frac{N\{t_j+\utilde{\lambda}_j' \utilde{\theta}\}}{\sqrt{1-\utilde{\lambda}_j' \utilde{P} \utilde{\lambda}_j}}\} \qquad [33]$$

in place of

$$\hat{y}_j = N\,\{f_{j0} + \utilde{f}_j' \utilde{\theta}\} \qquad [34]$$

(It should be noted that in the unidimensional case, Bock and Lieberman, 1970, employ both these parameterizations.)

The equivalence of Christoffersson's (1975) tetrachoric series to McDonald's (1967a) harmonic analysis has both theoretical and practical consequences, apart from the possibility that a few research workers interested in IRT but not in factor analysis may have missed the point that Christoffersson's (1975) "factor analysis of dichotomized variables" is indeed the random-regressors multidimensional normal ogive model with linear combination rule. Purely in terms of theory it is pleasing to find that the same result, Equations 27 and 30, can be obtained either as the solution to the problem of evaluating a double integral connected with the normal distribution or as a solution to the problem of approximating a strictly nonlinear regression function by a wide-sense linear regression function. The main practical implication of the result is that it gives grounds for reducng numerical work by using what at first sight would appear to be unacceptably crude approximations, such as the simple first-term approximation in Equation 21. That is, fitting an approximation to the model may be considered instead of fitting the model itself, thereby estimating the parameters reasonably precisely, where there might not have been an expectation of obtaining reasonably precise estimates of the parameters using very crude approximations to the integrals required by the model itself.

It may prove to be the case that any of the random-regressors methods discussed in this section can be expected to give reasonably satisfactory estimates of the item parameters from large samples if the assumption of a univariate or multivariate normal distribution of the latent trait is not too grossly violated. All of them allow the computation and inspection of residual covariances. It may be argued (McDonald, 1981, 1982b) that on the basis of the magnitudes of the residual covariances, a direct judgment of the adequacy of fit of the model (including the crucial question of the dimensionality of the data) is of more importance than a Neyman-Pearson statistical decision to reject or not reject the model at the given sample size, since as a restrictive statistical hypothesis the model is never correct in applications to real data. The use of the "heuristic" methods that do not yield a test of significance can thus be rationalized, at least for large data sets for which such a test is not available in practice.

Certainly one limitation of the random-regressors treatment is that imposed by the normal distribution assumption. It seems to be regarded as a more dangerous assumption for item response models than in its ubiquitous applications to linear common factor analysis and other models for the structural analysis of multivariate data. Another general limitation, briefly described in the introduction, is that random-regressors theory for a wide range of item characteristic curves is difficult to develop, while the corresponding fixed-regressors theory is generally straightforward. Even the logistic model, which has a number of mathematical properties that cause it to be preferred to the numerically almost indistinguishable normal ogive model, does not lend itself so readily to treatment as a random-regressors model. Also, both Christoffersson's (1975) "factor-analytic" theory and the multidimensional theory based on McDonald's (1967a) harmonic analysis yield a random-regressors treatment for the multidimensional normal ogive model with a linear combination rule for the $\underline{k}$ latent traits, but it is not at all obvious how to extend this treatment to allow nonlinear combination rules, as in the model:

$$\hat{y}_j = N\{f_{j0} + f_{j1}\theta_1 + f_{j2}\theta_2 + f_{j3}\theta_1\theta_2\} \qquad [35]$$

or

$$\hat{y}_j = N\{f_{j0} + f_{j1}\theta_1 + f_{j2}\theta_1^2\} . \qquad [36]$$

To see how such models might be dealt with the next section will be concerned with fixed-regressors theory.

Fixed-Regressors Theory

In this section no attempt will be made to review previous work. It must suffice to note that direct attacks on the problem of jointly estimating item parameters and examinee "abilities" in the normal ogive, logistic, and Rasch models by maximum likelihood have been made by Kolakowski and Bock (1970), Wingersky and Lord (1973), and Wright and his associates (Wright & Stone, 1979), at least. Although there is some evidence from monte carlo studies that these methods yield satisfactory estimates of the item parameters, it seems broadly true that they have not been made to yield a satisfactory test of fit, since the likelihood function has not been made to yield a loss function, measuring poorness of fit, and no statistics are computed to test for departure in the sample from obedience to the principle of local independence, given the hypothesized (unit) dimensionality. (Wright and Stone have computed fit statistics that are functions of the residuals of each item. These quantities do not bear any obvious relationship to the principle of local independence.) Consider next fixed-regressors theory for unidimensional and multidimensional latent trait models based directly on the nonlinear common factor model Equation 1, with the weak principle of local independence. The general model Equation 1 is rewritten as a fixed-regressors model,

$$y_{ji} = \phi_j(\theta_{1i}, \ldots, \theta_{ki} ; \beta_{j1}, \ldots, \beta_{js}) + e_{ji},$$

$$j = 1, \ldots, n \; ; \; i = 1, \ldots, N \, , \qquad [37]$$

where $\beta_{j1}, \ldots, \beta_{js}$ is a set of parameters describing the *j*th regression function. The residual covariances q_{jk} are defined by

$$q_{jk} = \frac{1}{N} \sum_i e_{ji} \, e_{ki} \, , \qquad [38]$$

and the (weak) principle of local independence is expressed as

$$\mathcal{E}\{q_{jk}\} = u_j^2 \; , \; j = k \, ,$$

$$= 0 \; , \; j \neq k \, . \qquad [39]$$

The n × n matrix of sample residual covariances is written $\underset{\sim}{Q} = [q_{jk}]$, and $U^2 = \mathrm{Diag}\left\{u_1^2, \ldots, u_n^2\right\}$, the expected value of $\underset{\sim}{Q}$ under the hypothesis.

McDonald (1979, in press) shows that any model of the type Equation 37 may be fit by choosing values of $\theta_{1i}, \ldots, \theta_{ki}, \beta_{j1}, \ldots, \beta_{js}, u_1^2, \ldots, u_n^2$, to make $\underset{\sim}{Q}$ as close as possible to a diagonal matrix in one of two senses. With respect to these parameters, either the loss function

$$\omega = \tfrac{1}{2}\mathrm{Tr}\{(\underset{\sim}{Q} - \underset{\sim}{U}^2)^2\} \, , \qquad [40]$$

or the loss function

$$\ell = -\tfrac{1}{2} \log \left| (\mathrm{Diag}\, \underset{\sim}{Q})^{-\frac{1}{2}} \underset{\sim}{Q} (\mathrm{Diag}\, \underset{\sim}{Q})^{-\frac{1}{2}} \right| \, . \qquad [41]$$

may be minimized. The first of these may be recognized as Harman's well-known MINRES loss function in linear common factor analysis, by the use of which there is an endeavor to minimize the squares of the residual covariances of distinct variables. The second corresponds to the maximum determinant method in linear factor analysis (Howe, 1955). By its use there is an endeavor to maximize the determinant of the correlation matrix of the residuals, thus making it as close as possible to an identity matrix. If the residuals can be assumed to have a multivariate normal distribution, which is certainly not the case for binary items, the second loss function also yields maximum likelihood ratio estimators (McDonald, 1979).

In principle, any prescribed model of the type Equation 37 can be fitted to binary items or test scores by minimizing either of the loss functions Equations 40 and 41; the goodness of fit of the model may be judged on the basis of the magnitudes of the loss functions. An inspection of the arrangement of magnitudes of the residual covariances in the residual covariance matrix may indicate the source of any disagreement between the model and the data and may suggest a revised model. To fit any model whatsoever of type Equation 37, it is a straightforward matter to program expressions for the loss function and its derivatives with respect to the parameters of the model and to apply a minimization algorithm, such as the method of conjugate directions, from a suitable starting point. It proves to be necessary to alternate between minimization of one of the loss functions Equation 40 or 41 with respect to the examinee parameters θ_{ji}, and minimization of the usual mean-square residual

$$\tau = \mathrm{Tr}\{\underset{\sim}{Q}\} \qquad [42]$$

with respect to the parameters of the regression on the current θ_{ji} values. It may be shown (see McDonald, in prep.) that the minimum point of τ with respect to the parameters of the regression is the same as the minimum of Equation 40 or 41 with respect to those parameters. Minimization of Equation 40 or 41 with respect to the parameters of the regression yields poor estimates in general. Curiously, the better the fit to the model, the poorer such estimates would be, and alternating the loss functions would be better.

Thus far, the only case of the model Equation 37 that has been extensively investigated is a polynomial model with pairwise interaction terms (Etezadi-Amoli & McDonald, in press; McDonald, 1979), i.e., a model of the type

$$y_{ji} = \sum_{p=1}^{k} \sum_{q=1}^{r_p} \beta_{jpq} \theta_p^{\,q} + \sum\sum_{p \neq p'} \gamma_{jpp'} \theta_p \theta_{p'} + e_{ji} , \qquad [43]$$

a wide-sense linear model which includes as a special case the simple polynomial model

$$y_{ji} = \sum_{q=1}^{r} \beta_{jq} \, \theta^q + e_{ji} . \qquad [44]$$

However, unlike random-regressors theory, the fixed-regressors theory just described should be just as easily applied to strictly nonlinear models, for example, to the multidimensional normal ogive model Equation 14 rewritten as

$$y_{ji} = N(f_{j0} + \underset{\sim}{f}'_j \underset{\sim}{\theta}_i) + e_{ji} , \qquad [45]$$

where, if the observed variables are binary item responses, the conditional distribution of y_{ji} given θ_i is binomial, with heteroscedastic variances given by

$$\mathcal{E}\{e_{ji}^2 \mid \theta_i\} = N(f_{j0} + f_j\theta_i) \quad [1 - N(f_{j0} + f_j'\theta_i)] \ , \qquad [46]$$

and certainly not normal. Again, as in the random-regressors form of this model, a pattern for the parameters f_j corresponding to simple structure may be prescribed or they may be constrained to be equal in columns of the matrix

$$\begin{bmatrix} f_1' \\ \vdots \\ f_n' \end{bmatrix} \qquad [47]$$

to yield a multidimensional counterpart of the Rasch model. In contrast to the random-regressors form of the model, any convenient and appropriate nonlinear function for N(·) may be substituted--the logistic function, for example--without creating any mathematical problems for the implementation of the theory. Perhaps more interestingly, the fixed-regressors version of the model allows escape from the restrictions of the linear combination rule. The model $N(f_{j0} + f_j'\theta_i)$ is a transformation, suitable for binary items, of the linear common factor model; and, as noted earlier, the regression functions are constant on planes orthogonal to f_j. Whereas the application of random-regressors theory to models such as Equation 35 and 36 seems to yield intractable mathematical problems, the fixed-regressors versions of such models

$$y_{ji} = N(f_{j0} + f_{j1}\theta_{1i} + f_{j2}\theta_{2i} + f_{j3}\theta_{1i}\theta_{2i}) + e_{ji} \qquad [48]$$

and

$$y_{ji} = N(f_{j0} + f_{j1}\theta_{1i} + f_{j2}\theta_{1i}^2) \qquad [49]$$

should be just as easy to fit as the model Equation 14, so the limitations of the linear combination rule are easily escaped.

A great deal of numerical work is needed on models arising out of the fixed-regressors treatment of Equations 1 and 2 by the weak principle of local independence before there can be confidence that the approach is generally practical. The virtues of this approach consist in its generality, its freedom from distribution assumptions, and the fact that it directly embodies the weak principle of local independence and thus tests the agreement of the chosen model, with prescribed dimensionality, with the data. An obvious disadvantage is that even with the use of an efficient minimization algorithm, such as the method of conjugate directions, limitations of sample size, which may be serious given the poor quality of information in binary data, are immediately apparent, whereas random-regressors methods are not limited by sample size.

A Numerical Example

The fixed-regressors methods described above have not yet been programmed for regression curves, unidimensional or multidimensional, that are suitable for binary responses. A program for the unidimensional normal ogive model based on McDonald's (1967a) harmonic analysis has been written by Fraser, who has recently extended this work to cover the multidimensional model Equation 14. A detailed account of the method used in these programs and of their numerical behavior will be given elsewhere (McDonald & Fraser, in prep.).

Here, a single numerical illustration of the multidimensional version of the method will be given to illustrate the general claim made above--that even the linear approximation can yield reasonable estimates because of the weighted least squares property of the polynomial series approximation to the normal ogive.

Let p_j be the proportion of examinees passing item j, and p_{jk} be the proportion passing items j and k, in a sample of size N. In program NOHARM (Normal Ogive Harmonic Analysis Robust Method) the threshold or position parameter is estimated in closed form by solving the sample analogue of Equation 26, i.e., by obtaining

$$\hat{t}_j = N^{-1}(p_j) \ , \qquad [50]$$

and the parameters $\underset{\sim}{f}'_j$ are obtained by ordinary least squares, minimizing

$$\phi = \sum_{j \neq k} (p_{jk} - \pi_{jk}^{(r)})^2 \ , \qquad [51]$$

where $\pi_{jk}^{(r)}$ is the r-term approximation to $\pi_{jk}^{(\infty)}$ in Equation 19, by a quasi-Newton algorithm. The combination of the random-regressors model and the weak principle of local independence with the use of ordinary least squares makes it possible to analyze quite large numbers of items with unlimited sample size. [With commonly available computer configurations, it should be possible to fit an equivalent of the Rasch model to 200 items, and the 2-parameter model (or the 3-parameter model with supplied pseudo-guessing parameters) to more than 100 items, with an unlimited number of examinees.

Data from sections of the LSAT have been used by a number of writers to illustrate IRT. In particular, LSAT7 has been treated by Christoffersson (1975) as a two-dimensional case of the normal ogive model. The sample raw product-moment matrix from LSAT7, with N = 1,000, is given in Table 1. Program NOHARM (McDonald & Fraser, in prep.) was used to fit model Equation 14 to the data, first using the linear approximation, then using the cubic approximation to the harmonic series. In both cases, orthogonal latent traits were prescribed, and an unrestricted 5 × 2 matrix of the parameters $\underset{\sim}{f}'_j$, $j = 1, \ldots, 5$, in Equation 14 was fitted and then rotated by varimax for comparison with Christoffersson's result. Table 2 gives the estimates from the linear and cubic approximation,

Table 1
Sample Raw Product-Moment Matrix for
Five Items from LSAT-7

Item	1	2	3	4	5
1	.828				
2	.567	.658			
3	.664	.560	.772		
4	.532	.428	.501	.606	
5	.718	.567	.672	.526	.843

after varimax rotation, in the parameterization Equation 14, while Table 3 gives the corresponding results after reparameterization by Equation 32, together with Christoffersson's (1975) generalized least squares estimates.

Table 2
Parameter Estimates from NOHARM

Item	Linear Approximation			Cubic Approximation		
	f_{j0}	f_{j1}	f_{j2}	f_{j0}	f_{j1}	f_{j2}
1	2.101	1.966	.253	1.975	1.819	.213
2	.491	.231	.636	.483	.231	.594
3	1.361	.493	1.446	1.270	.420	1.314
4	.295	.345	.289	.295	.341	.295
5	1.099	.373	.227	1.086	.331	.234

Table 3
Reparameterized Estimates

Item	Linear			Cubic			Christofferson		
	t_j	λ_{j1}	λ_{j2}	t_j	λ_{j1}	λ_{j2}	t_j	λ_{j1}	λ_{j2}
1	-.946	.886	.114	-.946	.872	.102	-.950	.796	.146
2	-.399	.191	.527	-.407	.195	.501	-.406	.197	.475
3	-.746	.270	.792	-.745	.246	.771	-.747	.212	.826
4	-.260	.315	.264	-.269	.311	.269	-.270	.325	.258
5	-.998	.342	.208	-1.007	.307	.213	-1.007	.310	.225

Table 4 gives the residual covariance matrices for the linear approximation and the cubic approximation. In this case, the results from the latter can hardly be claimed to be an improvement on those from the former, and both are in reasonable agreement with Christoffersson's results based on a 10-term series. The example seems to support the claim made earlier that the normal ogive model is remarkably well fitted by fitting the approximating linear model.

Table 4
Residual Covariance Matrices for Linear Approximation (Lower Triangle) and Cubic Approximation

Item	1	2	3	4	5
1		-.0006	.0002	.0002	.0002
2	-.0007		-.0001	-.0012	.0031
3	.0006	.0000		.0010	-.0020
4	.0001	-.0012	.0010		-.0007
5	.0000	.0031	-.0025	-.0001	

Conclusions

It has been suggested that there are a number of advantages to be obtained from treating item response theory as a special case of nonlinear common factor analysis. Nonlinear common factor analysis supplies a general yet rigorous definition of the dimensionality of a set of tests or of a set of dichotomous or polychotomous items, as the number of common factors or latent traits required in the model. This definition replaces ill-defined notions of homogeneity and internal consistency which have persisted in the literature, without clear explication, as near synonyms for unidimensionality (see McDonald, 1981). Nonlinear common factor analysis yields a natural embodiment of the weak principle of local independence in appropriate loss functions that can be used to fit item response models and to assess their adequacy as descriptions of the data. The recognition of latent traits as synonymous with common factors gives appropriate guidance for interpretation, as well as for estimating and testing the model, in both unidimensional and multidimensional versions. Note, for example, that Christoffersson's (1975) factor-analytic parameterization of the multidimensional normal ogive model, Equation 33, suggests the use of factor analytic standards for "salient" factor loadings in the identification of latent traits.

Within limitations set by the assumption of a normal distribution of latent traits, random-regressors theory enables the fitting of item parameters in the unidimensional and multidimensional normal ogive model with unlimited sample sizes and quite large item sets with very satisfactory numerical efficiency. There is usually little interest in the values of the latent traits of the examinees in the calibration sample, and if there is interest, these can easily be estimated in a second stage after the item parameters have been estimated. Random-regressors treatments do, then, have much to recommend them.

Fixed-regressors theory based on McDonald's (1979) treatment of common factor analysis is, in principle, applicable to any prescribed unidimensional or multidimensional model in item response theory, though so far it has been applied only to the rather inappropriate polynomial model. The advantages of this approach are (1) that it is not limited to the use of a linear combination rule for multiple latent traits and (2) that it easily permits the introduction of interaction terms. Applications of the fixed-regressors treatment are subject to limitations of sample size, which may prove severe in the case of binary data, where large sample sizes are generally desirable.

Another contribution of the factor analytic perspective on item response theory, not discussed in this paper, consists in the treatment of the invariance of item parameters across populations as closely related to the traditional treatments of factorial invariance (see McDonald, 1982b). Perhaps enough has been suggested to persuade researchers in the field of item response theory to recognize at least the theoretical value of seeing it as part of a more general psychometric unity.

REFERENCES

Bock, R. D., & Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-22.

Etezadi-Amoli, J., & McDonald, R. P. A second generation nonlinear factor analysis. Psychometrika, in press.

Howe, W. G. Some contributions to factor analysis (Report No. ONRL 1919). Oak Ridge TN: Oak Ridge National Laboratory, 1955.

Kolakowski, D., & Bock, R. D. A Fortran-IV program for maximum likelihood item analysis and test scoring: Normal ogive model (Research Memorandum No. 12). Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1970.

Lazarsfeld, P. F., & Henry, N. W. Latent structure analysis. Boston: Houghton-Mifflin, 1968.

Lord, F. M. A theory of test scores. Psychometric Monographs, 1952, No. 7.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

McDonald, R. P. A note of the derivation of the general latent class model. Psychometrika, 1962, 27, 203-206. (a)

McDonald, R. P. A general approach to nonlinear factor analysis. Psychometrika, 1962, 27, 397-415.

McDonald, R. P. Nonlinear factor analysis. Psychometric Monographs, 1967a, No. 15.

McDonald, R. P. Numerical methods for polynomial models in nonlinear factor analysis. Psychometrika, 1967, 32, 77-112.

McDonald, R. P. Factor interaction in nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology*, 1967, *20*, 209-215.

McDonald, R. P. *PROTEAN--A CDC computer program for nonlinear factor analysis* (Research Memorandum RM-67-26). Princeton NJ: Educational Testing Service, 1967. (d)

McDonald, R. P. A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 1978, *31*, 59-72.

McDonald, R. P. The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research*, 1979, *14*, 21-38.

McDonald, R. P. A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. *British Journal of Mathematical and Statistical Psychology*, 1980, *33*, 161-183.

McDonald, R. P. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 1981, *34*, 100-117.

McDonald, R. P. Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology*. Australian Council for Educational Research, 1982. (a)

McDonald, R. P. Linear versus nonlinear models in latent trait theory. *Applied Psychological Measurement*, 1982, *6*, 379-396. (b)

McDonald, R. P. Exploratory and confirmatory nonlinear factor analysis. In H. Wainer (ed.), *Festschrift for F. M. Lord*. Hillsdale NJ: Erlbaum, in press.

McDonald, R. P., & Ahlawat, K. S. Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 82-99.

Muthén, B. Contributions to factor analysis of dichotomous variables. *Psychometrika*, 1978, *43*, 551-560.

Neyman, J., & Scott, L. Consistent estimates based on partially consistent observations. *Econometrika*, 1948, *16*, 1-32.

Wingersky, S., & Lord, F. M. *A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses* (Research Memorandum RM-73-2). Princeton NJ: Educational Testing Service, 1973.

Wright, B. D., & Stone, M. H. *Best test design*. Chicago: Mesa Press, 1979.

APPENDIX

To obtain Equations 16 and 19 from Equation 14, the procedure is as follows: A transformation of $\underset{\sim}{\theta}$ into $\underset{\sim}{\theta}^*$ is sought, such that $\hat{y}_j$ is a function of one component of $\underset{\sim}{\theta}^*$, say the first, only. Define $\underset{\sim}{T}_j$ by

$$\underset{\sim}{T}_j = \frac{1}{\delta_j} \begin{bmatrix} \underset{\sim}{f}'_j \\ \underset{\sim}{f}'_{2j} \\ \vdots \\ \underset{\sim}{f}'_{kj} \end{bmatrix} \qquad [A1]$$

where

$$\delta_j = \underset{\sim}{f}'_j \underset{\sim}{f}_j \qquad [A2]$$

and $\underset{\sim}{f}_{\ell j}$ such that

$$\underset{\sim}{f}'_{\ell j} \underset{\sim}{f}_j = 0 \ , \qquad [A3]$$

$$\underset{\sim}{f}'_{\ell j} \underset{\sim}{f}_{mj} = 0 \ , \ \ell \neq m \ , \qquad [A4]$$

and

$$\underset{\sim}{f}'_{\ell j} \underset{\sim}{f}_{\ell j} = \delta_j^2 \ , \ \ell = 2, \ldots, k \ . \qquad [A5]$$

(That is, the rows of $\underset{\sim}{T}_j$ are completed arbitrarily after the first to yield an orthogonal matrix.)

Define

$$\underset{\sim}{\theta}^* = \frac{\delta_j}{d_j} \underset{\sim}{T}_j \underset{\sim}{\theta} \qquad [A6]$$

where

$$d_j^{\,2} = \underset{\sim}{f}'_j \underset{\sim}{P} \underset{\sim}{f}_j \qquad [A7]$$

so that, in particular,

$$\theta_1^* = \frac{1}{d_j} \underset{\sim}{f}_j{}' \underset{\sim}{\theta} \ , \qquad [A8]$$

with

$$\mathcal{E}\{\theta_1^*\} = 0 \ ; \ \mathcal{E}\{\theta_1^{*2}\} = 1 \qquad [A9]$$

(since $\mathcal{E}\left\{\mathbf{f}_j'\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{f}_j\right\} = \mathbf{f}_j'\mathbf{P}\mathbf{f}_j$).

The converse transformation to Equation A6 is

$$\boldsymbol{\theta} = \frac{d_j}{\delta_j}\mathbf{T}_j'\boldsymbol{\theta}^* , \qquad [A10]$$

since $\mathbf{T}_j$ is orthogonal. Then

$$N\,(f_{j0} + \mathbf{f}_j'\boldsymbol{\theta}) = N(f_{j0} + \frac{d_j}{\delta_j}\mathbf{f}_j'\mathbf{T}_j'\boldsymbol{\theta}^*) \; ; \qquad [A11]$$

and since, by construction of $\mathbf{T}_j$,

$$\mathbf{f}_j'\mathbf{T}_j' = [\delta_j \; 0 \; 0 \; \ldots \; 0] \qquad [A12]$$

gives, further,

$$N\,(f_{j0} + \mathbf{f}_j'\boldsymbol{\theta}) = N\,(f_{j0} + d_j\theta_1^*) \; . \qquad [A13]$$

Since $N(f_{j0} + d_j\theta_1^*)$ is a function of one variable only, by the properties of the unidimensional case, it may be represented by the infinite polynomial series

$$\phi_j^{(\infty)}\,(\theta_1^*) = \sum_{p=0}^{\infty} b_{jp}h_p(\theta_1^*) \qquad [A14]$$

where b_{j0}, b_{jp} are obtained by substituting d_j given by Equation A7 for f_{j1} in Equations 11 and 12. Equation 16 then follows.

Further, according to Lancaster (1958), if (x, y) are bivariate normally distributed with means zero, variances unity, and correlation ρ, then

$$\begin{aligned} \mathcal{E}\{h_p(x)h_q(y)\} &= \rho^p \; , \; p = q \\ &= 0 \; , \; p \neq q \; , \end{aligned} \qquad [A15]$$

where, as before, $h_p(\cdot)$ is the normalized Hermite-Tchebycheff polynomial given by Equation 10. It follows that

$$\pi_j = b_{j0} \qquad [A16]$$

and

$$\pi_{jk}^{(\infty)} = P\{y_j = 1, \; y_k = 1\} \; ,$$

$$= \mathcal{E}\{N(f_{j0} + \underset{\sim}{f}_j'\underset{\sim}{\theta})\; N(f_{k0} + \underset{\sim}{f}_k'\underset{\sim}{\theta})\},$$

$$= \sum_{p=0}^{\infty} b_{jp} b_{kp} \left(\frac{1}{d_j} \underset{\sim}{f}_j' \underset{\sim}{P} \underset{\sim}{f}_k \frac{1}{d_k}\right)^{p} , \qquad [A17]$$

as stated in Equation 19.

DISCUSSION

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE

I found this paper interesting, particularly the correspondence between McDonald's work and Christofferson's. Since the approximation by polynomials to many different functions is quite feasible (as I have often done in different contexts in my own research), I consider McDonald's use of the polynomial model in nonlinear factor analysis appropriate. His assertion that latent trait theory is a special case of nonlinear factor analysis, recognizing latent traits as synonymous with common factors, does not quite convince me, however.

It should be noted that factor analysis is basically a population-restricted technique in the sense that the result describes a specific population of individuals, starting from its correlation or covariance matrix. This is still true even if simultaneous factor analysis (in which two or more populations are simultaneously dealt with) is considered. In contrast, latent trait theory is basically a population-free theory. Research on the estimation of the operating characteristics without assuming any mathematical form--such as Lord's, Levine's, and mine--do not assume any specific distribution for the ability distribution. To give an example, if normality is assumed for the latent trait distribution in my Conditional P.D.F. Approach, only the Normal Approach Method is needed; the reason why I also developed the Pearson System Method and the Two-Parameter Beta Method is to provide methods applicable when the latent trait distribution cannot be approximated by a normal distribution.

For this reason and others, my conclusion is that latent trait theory is a more comprehensive theory, and it is more appropriate to say that factor analysis is a special case of latent trait theory. There have been, at least, several theorists in latent trait theory who are aware of the similarities between the two theories. Lord (1952) used factor analysis as a tool in defining the unidimensional latent space as the dominating first common factor. I (Samejima, 1974) discussed the similarities of the two theories. Although it is true that many researchers who either work on or apply latent trait theory are unaware of the similarities, the situation is probably better compared with the situation of the factor analysts. After all, how many factor analysts are aware of the similarities between factor analysis and latent trait theory and have tried to gain benefits from research on latent trait theory?

In fixed regressors theory, there are some problems in using the maximum likelihood estimate (MLE) of the common factor, or the latent trait, when any regression is considered. Although under certain conditions, the expectation of the MLE approximately equals the expectation of the original variable, the mth moment about the mean is given by

$$E[\hat{\theta}_V - E(\hat{\theta}_V)]^m = \sum_{r=0}^{m} \binom{m}{r} E[\{\theta - E(\theta)\}^{m-r} E\{(\hat{\theta}_V - \theta)^r|\theta\}] \qquad [1]$$

where θ is the latent trait and $\hat{\theta}_V$ is its MLE based upon the response pattern V. When $m = 2$, for example, the above equation gives the variance of $\hat{\theta}_V$, which, in many cases, is greater than the variance of θ. Generally speaking, the regression of θ on $\hat{\theta}_V$ is not even linear. Thus, some adjustment, at least, is needed when $\hat{\theta}_V$ is used as the substitute of θ in obtaining any regression to make up for the uneven stretches.

REFERENCES

Lord, F. M. A theory of test scores. Psychometric Monographs, 1952, No. 7.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39, 111-121.

SOME LATENT TRAIT THEORY IN A MULTIDIMENSIONAL LATENT SPACE

MARK D. RECKASE AND ROBERT L. MCKINLEY
THE AMERICAN COLLEGE TESTING PROGRAM

Measuring instruments that are used to determine an individual's level of performance on a psychological or educational trait are seldom truly unidimensional. Certainly, tests based on number series or vocabulary knowledge may approximate unidimensional measures, but even these narrowly focused tests usually measure more than one trait (see, e.g., Holzman, Glaser, & Pellegrino, 1980).

Alternatively, many tests are purposely designed to measure more than one trait. The English Usage Test from the ACT Assessment battery, for example, measures skills in punctuation, grammar, sentence structure, diction and style, and logic and organization (American College Testing Program, 1980). These topics have been included in the test in order to assess more thoroughly the skills acquired in high school English than could be obtained from a unitary measure. There is also a statistical motivation for constructing a test that measures more than one dimension. To maximize a test's ability to predict a criterion measure, the test should have items that have a high correlation with the criterion but low intercorrelations among themselves (Lord & Novick, 1968). Following this selection rule results in a test that measures many dimensions.

The fact that measuring devices seldom measure single dimensions has serious consequences for the application of item response theory (IRT) to test data. A basic assumption of most of the IRT models currently being applied is that the measuring instrument measures a single trait (Lord, 1980; Lord & Novick, 1968). To the extent that this assumption is violated, these IRT models may not be appropriate. Since tests seldom measure single dimensions, the unidimensional IRT models are only applicable if they can be shown to be robust to the violation of the unidimensionality assumption or if the items in a test can be sorted into subtests that measure a single dimension.

The issue of the robustness of two IRT models--the 1-parameter and 3-parameter logistic models--to violations of the unidimensionality assumption has been addressed by Reckase (1977). He found that even when the proportion of variance accounted for by the dominant dimension was as low as 20%, the two models still resulted in reasonable ability estimates. However, since these estimates were of the dominant dimension, much information was lost about other traits being measured by the other dimensions in the tests. On the other hand, if the measuring instrument in question measured several traits with equal emphasis, the meaning of the ability estimates was difficult to define. Thus, although the models do seem to be somewhat robust to violations of the unidimensionality assumption, it is at the cost of lost information or poorly defined traits. It would seem that a better approach would be to estimate the ability on each dimension separately.

Two different alternatives exist for obtaining estimates of the traits measured by a test when more than one trait is being measured. First, the items in the test may be subdivided into groups of items that are sensitive to differences on one of the dimensions. This procedure breaks down the test into a series of unidimensional subtests. Unfortunately, no procedure exists that adequately performs this function when dichotomously scored test items are used (Reckase, 1981). Factor analysis is the procedure most commonly used for sorting items, but factor analysis suffers from several problems due to the choice of the correlation coefficient, the effects of guessing, and the determination of the number of factors (Kim & Mueller, 1978). Therefore, in many cases the formation of unidimensional subsets of items is not a reasonable approach.

The second possible approach for obtaining estimates of the abilities on each dimension is to develop a multidimensional model of performance that relates dichotomous item responses to the magnitude of ability on each trait. Several models of this type have been presented in the IRT literature (Bock & Aitkin, 1981; Mulaik, 1972; Rasch, 1961; Samejima, 1974; Whitely, 1980), but little work has been done using these models in an applied testing setting. In fact, little research has been done to determine the characteristics of these models or the properties of the ability estimates obtained through their use.

From this discussion it should be evident that obtaining estimates of trait levels from a test that measures more than one trait is a difficult problem. Traditional models such as factor analysis and nonmetric multidimensional scaling are not well suited for use with dichotomously scored test items, and most of the IRT models that are designed for use with dichotomous test data assume a unidimensional test. The use of multidimensional IRT models may be the solution to this problem; however, little work has been done to demonstrate their usefulness. The purpose of this paper is to review the existing multidimensional IRT models and to show how one of the models can be applied to the estimation of abilities from a test measuring more than one dimension.

Definition of the Problem

Most of the IRT models currently in use assume that the test being analyzed measures a unidimensional latent trait. This means that all persons having the same amount of the trait, θ, should have the same probability of a correct response to a dichotomously scored item. If individuals with the same level of the single trait have different probabilities of a correct response to a test item, this implies that at least one other trait is involved in responding to the item. If only two dimensions are required in the solution of the item, then all persons that have the same values on these two dimensions, $[\theta_1, \theta_2]$, should have the same probability of a correct response. Again, if the examinees have different probabilities of a correct response, at least one more dimension is indicated. Once the number of dimensions, $\underline{n}$, is determined that results in a constant probability of a correct response for all persons with the same set of abilities, $\theta_1, \theta_2, \ldots, \theta_n$, the size of the complete latent space has been de-

BEST COPY AVAILABLE

fined. This concept is discussed in more detail by Lord and Novick (1968).

Note that this method of defining the size of the complete latent space emphasizes the ability dimensions of the examinees while treating the test item as a constant stimulus. No information is given concerning the characteristics of the items. In order to determine the characteristics of the test items, critical features of the surface describing the relationship between the probability of a correct response and a person's position in the θ space must be defined. Two such features that are typically used in IRT are (1) the difficulty of the item (location of the point of inflection of the item characteristic curve, or ICC) and (2) the discriminating power of the item (related to the slope of the ICC at the point of inflection). If the relationship between the probability of a correct response and a person's location in the θ space can be described by a sufficiently well-behaved mathematical function (e.g., the logistic function), the concept of difficulty and discrimination can be generalized to items that measure more than one dimension in the complete latent space.

Suppose that the relationship between the probability of a correct response to a dichotomously scored test item and a person's location in the θ space is given by a function that is monotonically increasing for all θ dimensions and is asymptotic to 0 and 1 as each $\theta \to -\infty$ and $\theta \to \infty$, respectively. That is,

$$f(\theta_{ij}) < f(\theta_{ik}) \qquad i = 1, \ldots, n , \qquad [1]$$

if $\theta_i < \theta_{ik}$ for all i, j, and k, where i indicates the dimension and j and k indicate the person, and

$f(\theta_{ij}) \to 0$

as $\theta_{ij} \to -\infty$,

$f(\theta_{ij}) \to 1$

as $\theta_{ij} \to \infty$, and

$1 > f(\theta_{ij}) > 0$ for all i, j .

Then, the difficulty of the item can be defined as the values of $\underset{\sim}{\theta}$ for which

$$\frac{d^2 f(\underset{\sim}{\theta})}{d\underset{\sim}{\theta}^2} = 0 , \qquad [2]$$

if certain regularity conditions hold. This is the multivariate equivalent to the point of inflection of the univariate ICC.

For some functions, $f(\underset{\sim}{\theta})$, the second derivative will be undefined; and for others, Equation 2 will yield multiple solutions. However, for a class of models based on the logistic function, Equation 2 gives a solution that defines a difficulty function rather than a difficulty value for an item. This function is the locus of points in the $\underset{\sim}{\theta}$ space that yields a .5 probability of a correct response to the item. An example of the difficulty function for an item that measures two dimensions may help clarify this concept.

Suppose that the complete latent space is defined by two dimensions, θ_1 and θ_2, and that the relationship between the probability of a correct response to an item and the values of $\underset{\sim}{\theta}$ are given by the following function:

$$P(x = 1|\underset{\sim}{\theta}, \underset{\sim}{\sigma}) = \frac{e^{(\sigma_1\theta_1 + \sigma_2\theta_2 + \sigma_3)}}{1 + e^{(\sigma_1\theta_1 + \sigma_2\theta_2 + \sigma_3)}} , \qquad [3]$$

where σ_1, σ_2, and σ_3 are values related to the shape of the probability surface for this particular item. An example of the probability surface is given for $\sigma_1 = 1.5$, $\sigma_2 = .5$, and $\sigma_3 = .65$ in Figure 1. The difficulty function for this item is defined as

$$\frac{\delta^2 P(x = 1|\underset{\sim}{\theta}, \underset{\sim}{\sigma})}{\delta\theta_1^2} = 0 . \qquad [4]$$

The second derivative is only taken with respect to θ_1 in this case because the points of inflection define the same function in both dimensions. If for simplicity, P is used in place of $P(x = 1 | \underset{\sim}{\theta}, \underset{\sim}{\sigma})$, the second derivative is equal to

$$\frac{\delta^2 P}{\delta\theta_1^2} = \sigma_1{}^2 P(1 - 3P + 2P^2) . \qquad [5]$$

If this expression is set equal to zero and solved for P, three solutions result--0, .5, and 1. Since 0 and 1 are degenerate cases where $\underset{\sim}{\theta} = \pm\infty$, the difficulty function is defined as the intercept of the probability surface with the .5 plane (a plane parallel to the θ plane at P = .5)

The line of intersection of the .5 plane with the probability surface can be obtained by determining which values of θ result in a .5 probability of a correct response. Since the exponent in Equation 3 must be equal to zero to obtain a probability of .5, the appropriate values of $\underset{\sim}{\theta}$ are the solutions to the equation

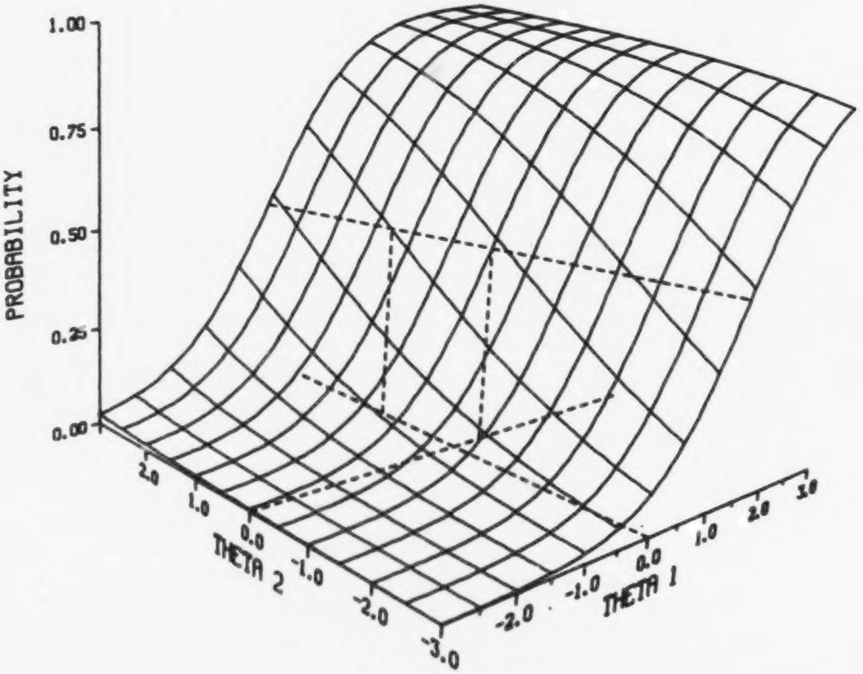
$$\sigma_1\theta_1 + \sigma_2\theta_2 + \sigma_3 = 0 . \qquad [6]$$

This is the equation of a straight line in the $\underset{\sim}{\theta}$ plane. In the usual linear form, the equation becomes

$$\theta_2 = -\frac{\sigma_1}{\sigma_2}\theta_1 - \frac{\sigma_3}{\sigma_2} . \qquad [7]$$

This line is shown as a dashed line on the .5 plane in Figure 1. Thus, for this example, the difficulty of the item is defined as a linear function instead of a single value.

Figure 1
An Example of a Two-Dimensional Probability Surface



The difficulty of the item in the usual latent trait sense can be determined on each dimension by holding the ability on the other dimension constant; for example, the point on the θ_2 scale that yields a .5 probability of a correct response when $\theta_1 = 0$ is $-\sigma_3/\sigma_2$, which is equal to $-(-.650/1.5) = .43$ for the surface given in Figure 1. Similarly, when $\theta_2 = 0$, $\theta_1 = -\sigma_3/\sigma_1 = -(-.650/.5) = 1.3$. Note that these "conditional difficulties" are different for each dimension even though there was only one summative term in the model, σ_3. When the dimensionality of the latent space is greater than two, the difficulty function for an item defines a hyperplane if a logistic model is used to describe the probability surface.

The definition of the discriminating power of the item in a multidimensional space can also be generalized from that used in the unidimensional case. In the unidimensional case, the discriminating power of an item is a function of the slope of the item response function at the point of inflection. The discriminating power of the item in a multidimensional space can likewise be determined by calculating the slope of the item response surface along the line defining the difficulty of the item.

This slope can be determined by evaluating the first derivative of the item response surface for values on the difficulty line. In the example given above, the first derivative of the item response surface with respect to θ_1 is

$$\frac{\partial P}{\partial \theta_1} = \sigma_1 P(1 - P) = \sigma_1 \left[\frac{e^{(\sigma_1\theta_1 + \sigma_2\theta_2 + \sigma_3)}}{1 + e^{(\sigma_1\theta_1 + \sigma_2\theta_2 + \sigma_3)}}\right]\left[\frac{1}{1 + e^{(\sigma_1\theta_1 + \sigma_2\theta_2 + \sigma_3)}}\right] . \quad [8]$$

If this function is evaluated for points on the line $\theta_2 = (-\sigma_1/\sigma_2)\theta_1 - \sigma_3$, the result is $.25\sigma_1$ for all values of θ_1. Thus, σ_1 is a discrimination parameter for the first dimension. Since the model is symmetric with respect to the θ's, the derivative with respect to θ_2 results in $.25\sigma_2$ when evaluated at all values on the difficulty line. Thus, both the difficulty function and the discrimination parameters are defined by the σ parameters in the exponent of this model. As with the difficulty function, the discrimination of a multidimensional item can be generalized to many dimensions. In the general case, the discrimination with respect to a dimension is the slope of the item response surface at the difficulty hyperplane. This slope may be a function of the $\underset{\sim}{\theta}$ vector in more complex models.

Up to this point, the complete latent space has been defined and the concept of an item response surface (IRS) has been introduced. Extensions of the unidimensional concepts of difficulty and discrimination have also been defined for the multidimensional IRS. The goal of this research, however, was to estimate the amount of ability an examinee possesses on each of these dimensions in the multidimensional latent space. Before this goal can be attained, two steps must be completed. First, a reasonable and convenient form for the IRS must be selected; and secondly, the parameters of the item response surface for each item must be determined.

Multidimensional Latent Trait Models

A number of models already exist in the literature for approximating the IRS. Each of these models will now be described, and the characteristics of the surface that they define will be summarized. Subsequently, one model will be selected for further analysis, leading to the estimation of model parameters.

Rasch's Model

The first of the models to be produced for approximating the IRS in a multidimensional space was presented by Rasch (1961). Although the model was not specifically designed to represent multidimensional data, Rasch indicated that vectors could be used for item and person parameters, thus extending the model to the multidimensional case.

The general form of the model is given by

$$P(x = i|\underset{\sim}{\sigma},\underset{\sim}{\theta}) = \frac{e^{\left(\sum_{\ell}^{n} w_{i\ell}\theta_\ell + \sum_{\ell}^{n} u_{i\ell}\sigma_\ell + \sum_{\ell}^{n}\sum_{m}^{n} v_{i\ell m}\theta_\ell\sigma_m + z_i\right)}}{\sum_{i}^{k} e^{\left(\sum_{\ell}^{n} w_{i\ell}\theta_\ell + \sum_{\ell}^{n} u_{i\ell}\sigma_\ell + \sum_{\ell}^{n}\sum_{m}^{n} v_{i\ell m}\theta_\ell\sigma_m + z_i\right)}} , \quad [9]$$

where

x is one of $i = 1, \ldots, k$ item responses;
θ is the person parameter vector with elements θ_ℓ, $\ell = 1, \ldots, n$;
σ is the item parameter vector with elements σ_ℓ, $\ell = 1, \ldots, n$;
w, u, and v are weights for the person and item dimensions;
z is a scaling constant for the item responses; and
n is the number of dimensions.

The model can be shown more conveniently in vector form as

$$P(x = i|\sigma,\theta) = \frac{e^{(W_i'\theta + U_i'\sigma + \theta' V_i \sigma + z_i)}}{\sum_i^k e^{(W_i'\theta + U_i'\sigma + \theta' V_i \sigma + z_i)}}\ , \qquad [10]$$

where W and U are vectors of weights for each item response, and V is a matrix of weights.

This model is extremely general, allowing both dichotomous and polychotomous scoring and containing both the 1-parameter and 2-parameter logistic models as special cases. Because the model is so general, it is difficult to determine the form of the item difficulty and discrimination functions. However, for the special case of a dichtomously scored item with W and U equal to unit vectors for a correct response and zero vectors for an incorrect response, V equal to the identity matrix for a correct response and a zero matrix for an incorrect response, and z equal to zero for all responses, the difficulty function is a hyperplane and the conditional slopes of the surface where it intersects the .5 plane are functions of σ_i. The model presented in Equation 3 and shown in Figure 1 is a special case of the general model.

Only one study is known that uses this general model to represent multidimensional item response data (Reckase, 1972), although there have been other applications of the model (Andersen, 1982; Andrich, 1978). In the Reckase (1972) study an attempt was made to estimate the parameters using a least squares procedure for a special case of this model, where V_i is a zero matrix for all responses and z_i takes on zero values for all responses. The attempt was not entirely successful, however, in that the fit of the multivariate model to multivariate data was no better than the fit of the simple Rasch (1960) model to the same data when estimates of the parameters were used. The poor results were attributed to two factors: (1) both the parameter vectors and the weights were estimated from the data and (2) the sample size used was too small to estimate accurately the large number of unknown variables. The parameter estimates were interpretable, however, suggesting that a less ambitious approach might be fruitful.

Mulaik's Model

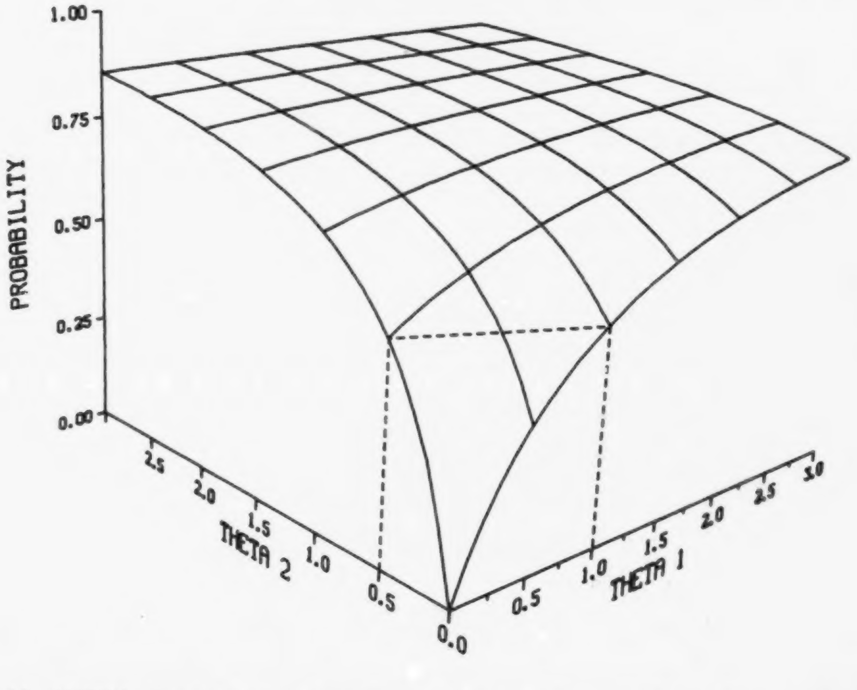
Another multidimensional model that was developed as an extension of the work of Rasch (1960) was proposed by Mulaik (1972). This model is given by

$$P(x = 1|\underline{\gamma},\underline{\pi}) = \frac{\sum_{i=1}^{n} \gamma_i \pi_i}{1 + \sum_{i=1}^{n} \gamma_i \pi_i} , \qquad [11]$$

where the γ_i's are item parameters and the π_i's are person parameters for the interaction of a person and an item in an *n*-dimensional space.

The previous definitions of the item difficulty and discrimination do not apply to this model, since the surface defined by Equation 11 does not have a point or line of inflection. However, the intersection of the surface with the .5 plane is a hyperplane and could be used to define item difficulty. Unlike the previous model, the conditional slope of the IRS at the intersection with the .5 plane is not simply a function of the item parameters but also depends on the ability parameters. A two-dimensional example of the response surface described by this model is presented in Figure 2.

Figure 2
Item Response Surface for Mulaik's Model



Mulaik (1972) presented a maximum likelihood procedure for estimating the parameters of this model, but it appears that it has not ever been applied. He cautioned that the amount of computation and the constraints required to estimate the parameters may be too great for the current generation of computers.

Sympson's Model

A third model that has been developed to describe the interaction of a person and an item in a multidimensional latent space was described by Sympson (1978). Rather than extend the 1-parameter logistic model, as done by Rasch (1961) and Mulaik (1972), Sympson (1978) based his model on an extension of the 3-parameter logistic model (Birnbaum, 1968). The mathematical expression for this model is given by

$$P(x = 1|\underset{\sim}{\theta},\underset{\sim}{a},\underset{\sim}{b},c) = c + (1 - c) \prod_{\ell=1}^{n} \left[1 + e^{[-1.7a_\ell(\theta_\ell - b_\ell)]}\right]^{-1}, \qquad [12]$$

where

$\underline{x}$ is the item response,
$\underset{\sim}{\theta}$ is a vector of ability parameters,
$\underset{\sim}{a}$ is a vector of discrimination parameters,
$\underset{\sim}{b}$ is a vector of difficulty parameters, and
$\underline{c}$ is a pseudo-chance level parameter.

An example of the surface defined by Equation 12 is given in Figure 3 for the two-dimensional case with parameters $c = .2$, $a_1 = .7$, $a_2 = 1.2$, $b_1 = -.6$, and $b_2 = .5$.

Unlike the models presented by Rasch (1961) and Mulaik (1972), the root of the second derivative of this equation does not define a difficulty function but gives a single value for each dimension. This value is simply the $\underline{b}$ parameter for that dimension. The difficulty of an item using this model can therefore be defined as the vector of $\underline{b}$ values, which defines a point in the multidimensional space.

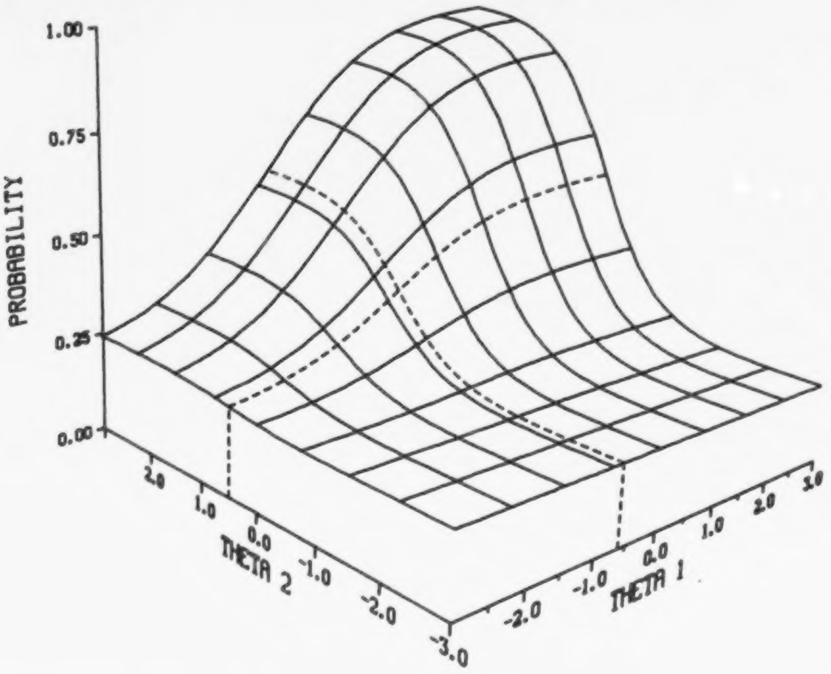
The slope of the IRS at the point of inflection for Sympson's (1978) model is given by

$$\frac{(1-c)(1.7)^n \prod_{\ell=1}^{n} a_\ell}{4^n},$$

which is solely a function of the item discrimination parameters and the pseudo-chance level parameter. If the slope of the function is determined at the difficulty point with respect to just one dimension, θ_1, the result is $(1 - c)(1.7)a_1/4 \times 2^{(n-1)}$, where $\underline{n}$ is the number of dimensions. Thus for this model the $\underline{a}$ vector defines the discrimination power of the item.

Sympson (1978) has done some preliminary work on estimating the parameters of this model for some simple cases, but no procedure has yet been published for the full multidimensional case. Lord (1978), in discussing Sympson's (1978)

Figure 3
Item Response Surface for Sympson's Model



paper, has suggested that a Bayesian or maximum likelihood procedure might be more fruitful than the method Sympson proposed. However, these methods have not been developed to the point where this model can be applied to actual test data.

Bock and Aitkin's Model

Bock and Aitkin (1981) suggested a multidimensional latent trait model that is an extension of the 2-parameter normal ogive model (Lord & Novick, 1968). They also indicated that this model is similar to the factor analysis procedures developed for dichotomous data by Christoffersson (1975) and Muthén (1978).

The mathematical form of the Bock and Aitkin (1981) model is given by the equation

$$P(x = 1 \mid \underset{\sim}{\theta}, \underset{\sim}{\sigma}, c) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{z(\underset{\sim}{\theta})} e^{\left(-\frac{t^2}{2}\right)} dt \,, \qquad [13]$$

where

$z(\underset{\sim}{\theta}) = c + \Sigma_{i=1}^{n} a_i \theta_i$,
$\underset{\sim}{\theta}$ is a vector of ability parameters,
$\underset{\sim}{a}$ is a vector of discrimination parameters, and
$\underline{c}$ is the item difficulty parameter.

Due to the similarity of the cumulative model and logistic functions, this model is very similar to one of the special cases of the general Rasch model given in Equation 9. Specifically, it corresponds to the case where the exponent of that model is equal to $\sum_i \sigma_i \theta_i + \sigma_{n+1}$ for a correct response and 0 for an incorrect response.

As with the general Rasch model, the difficulty function for the model is defined by the equation for a hyperplane

$$\Sigma a_i \theta_i + c = 0 \; . \qquad [14]$$

The conditional slope of the IRS for points on the difficulty function is given by $a_i/\sqrt{2\pi}$, demonstrating that the $\underline{a}$ vector is related to the discriminating power of the item. Since the two-dimensional IRS for this model is indistinguishable from the logistic surface presented in Figure 1, an example is not presented for this model.

Bock and Aitkin (1981) have produced an estimation procedure for their model based on the marginal maximum likelihood technique. They have applied this procedure to data assuming a two-dimensional solution. The results of the analysis showed that different quadrature procedures used in conjunction with marginal maximum likelihood techniques gave slightly different results but that a two-dimensional solution seemed to fit the data fairly well. No other applications of this model are known.

Samejima's Model

Samejima (1974) presented a more general version of the model suggested by Bock and Aitkin (1981) as a special case of her continuous response model in a multidimensional latent space. Her model is given by the equation

$$P^*_{z_g}(\underset{\sim}{\theta}) = \int_{-\infty}^{\underset{\sim}{a}_g'(\underset{\sim}{\theta} - \underset{\sim}{b}_g)} \psi_g(u)\,du \; , \qquad [15]$$

where

z_g is the point of dichotomy of the continuous trait measured by this item,
$P^*_{z_g}(\underset{\sim}{\theta})$ is the probability of a correct response,
$\underset{\sim}{a}_g$ is a vector of discrimination parameters,
$\underset{\sim}{\theta}_g$ is a vector of ability parameters,
$\underset{\sim}{b}_g$ is a vector of difficulty parameters, and
$\psi_g(u)$ is a twice differentiable function.

When ψ_g is the normal density function, Equation 15 is identical to Equation 13 with $c = -\Sigma a_i b_i$. When ψ_g is defined as the logistic density function, the model is a special case of the general Rasch model given in Equation 9. Samejima also points out the similarity of the model to linear factor analysis.

Whitely's Models

One other class of multidimensional latent trait models exists in the literature, but this class of models was developed from a different perspective than the others. The models presented up to this point generally consider the dimensions required to determine the complete latent space as unknown hypothetical constructs, the properties of which need to be discovered. In contrast, the class of the models proposed by Whitely (1980) considers the dimensions to be components in a cognitive model of performance. These dimensions are defined in advance of being estimated as particular cognitive processes.

The mathematical form of the model used by Whitely is similar to that used by Sympson (1981) in that it is composed of the product of separate logistic model terms. The particular model, called the multicomponent latent trait model, is given by the equation

$$P(x = 1|\underset{\sim}{\theta},\underset{\sim}{b}) = \prod_{i=1}^{n} \frac{e^{(\theta_i - b_i)}}{1 + e^{(\theta_i - b_i)}} , \qquad [16]$$

where the variables are as defined above. This model is the same as the Sympson (1978) model with $c = 0$ and $a_i = 1/1.7$ for all $\underline{i}$. Since it is a special case of the Sympson model, it also has a difficulty function equal to the $\underline{b}$ vector, and the slope at the point defined by the $\underline{b}$ vector is $1/4^n$. Whitely (1980) has also produced more complex versions of this model, but they are all composed of combinations of the 1-parameter logistic model.

Whitely (1980) has developed procedures for estimating the parameters of this model, but the estimation has been performed in a different manner than the other models that have been described. Whereas the estimation procedures for the other models have attempted to estimate the vector parameters from the dichotomous responses to test items, Whitely has developed an experimental design for collecting responses on each cognitive component separately. The parameters of each of the product terms of the model are then estimated separately using procedures developed for the unidimensional Rasch model. No descriptions of a procedure for simultaneous estimation of all of the parameters of the model has been found in the literature.

Comparison of the Multidimensional Models

An analysis of the six models that have been described above indicated that these models fall into three basic classes. The first of these classes (Class I) is of the form

$$P(x = 1|\underset{\sim}{\theta},\underset{\sim}{\sigma}) = \int_{-\infty}^{z(\underset{\sim}{\theta})} \psi(u)du , \qquad [17]$$

where $z(\underset{\sim}{\theta})$ is a linear function of the elements of $\underset{\sim}{\theta}$. The general Rasch (1961) model, the Bock and Aitkin (1981) model, and the Samejima (1974) model fall into this class. All of these models allow high ability on one dimension to compensate for low ability on another dimension, resulting in what Sympson (1978) has labeled as compensatory models. All of these models have linear difficulty functions when used with dichotomously scored data and have conditional slopes at points on the difficulty function that are functions of the corresponding discrimination parameters. These models are fairly simple from a mathematical point of view (especially when $\psi(u)$ is the logistic density function), and estimation procedures have been developed for the parameters (Bock & Aitkin, 1981).

Class II models contain a single example, the model proposed by Mulaik (1972). This model is of the form

$$P(x = 1|\underset{\sim}{\gamma},\underset{\sim}{\pi}) = \frac{\sum_{i=1}^{n} \gamma_i \pi_i}{1 + \sum_{i=1}^{n} \gamma_i \pi_i} , \qquad [18]$$

where the variables are as defined earlier. This model is also compensatory in Sympson's (1978) sense, but it is unlike the previous class in that the ability metric is defined from 0 to +∞ instead of from −∞ to +∞. This results in an IRS that does not have a point or line of inflection. If the difficulty of the item is defined by the intersection of the item with the .5 plane, the result is a linear function similar to that for the Class I models. The slope of the IRS at the difficulty function has the property of changing with its position relative to the ability dimensions. Mulaik (1972) has proposed an estimation procedure for this model, but there have been no studies to determine its practicality.

Class III models contain the models proposed by Sympson (1978) and Whitely (1980). These models take the form

$$P(x = 1|\underset{\sim}{\theta},\underset{\sim}{\sigma}) = \sigma_1 + (1 - \sigma_1) \prod_{i=1}^{n} P'_i(\theta_i) , \qquad [19]$$

where σ_1 is a lower asymptote parameter and $P'_i(\theta_i)$ is the probability of response with respect to a specific dimension. In Whitely's (1980) model the dimensions are defined as specific cognitive processes required to solve the problem proposed in the item, and the σ_1 parameter is assumed to be zero. In Sympson's (1978) model the dimensions are hypothetical traits based on commonalities among items.

Unlike the previous models, the Class III noncompensatory models do not allow a high ability on one dimension to compensate for a lower ability on another. The lowest of the values of $P'_i(\theta_i)$ defines the upper bound of $P(x = 1|\underset{\sim}{\theta},\underset{\sim}{\sigma})$. Although some work has been done on the estimation of parameters for the Class III models, no generally accepted algorithm for estimation of the parameters of these models is known to exist.

Several issues need to be considered in selecting one of these models as a description of the interaction between a person and an item. The first is whether the model is realistic. This depends on whether a compensatory or noncompensatory model is appropriate for actual persons and items. Unfortunately, this is a question that still needs to be answered, based on research in cognitive psychology. Although sufficient information is presently not available, the applicability of the models to actual testing situations may provide an answer.

Beyond questions of the psychological meaningfulness of the models are questions of practicality. The most well developed estimation procedures are available for the Class I models; and these models tend to have the most flexible options due to the characteristics of the exponential term. As a consequence, the Class I models tend to be more promising than the other models. Of the Class I models, the generalized Rasch model has the greatest flexibility in its options and is the most mathematically tractable. The remainder of this paper will therefore concentrate on the properties of this model and the procedures for the estimation of its parameters.

Application of the General Rasch Model

Although the general Rasch model is a generalization of the 1-parameter logistic model, a very simple model, in its most general form the model is very complex. A study (McKinley & Reckase, 1982) was thus undertaken to determine whether a less complex formulation of the model would be adequate for modeling multidimensional response data.

Method

Design. The general design of this study was to first evaluate the properties of a simple formulation of the general model and then to evaluate increasingly more complex versions of the model produced by inserting additional terms. The initial form of the model investigated is given by Equation 20:

$$P(x \mid \underset{\sim}{\theta}_j, \underset{\sim}{\sigma}_i) = \frac{1}{\gamma(\underset{\sim}{\theta}_j, \underset{\sim}{\sigma}_i)} \exp(\underset{\sim}{U}'\underset{\sim}{\sigma}_i + \underset{\sim}{W}'\underset{\sim}{\theta}_j) \ . \qquad [20]$$

For each level of model complexity, the properties of the model were investigated and the reasonableness and usefulness of the model were explored. This was done primarily by generating simulated test data to fit the particular form of the model being investigated and by analyzing that data in an attempt to assess how well the characteristics of the data matched the characteristics of real test data. If it were found that a particular form of the model could not be used to generate realistic data in terms of either dimensionality or item characteristics, then that form of the model was rejected and a different form of the model was investigated. Distinct special cases of the model were obtained by eliminating different terms from the general model by setting the appropriate parameter weights equal to zero.

Analyses. The analyses of the generated data that were performed included factor analysis and traditional item analysis. The purposes of the analyses were three-fold. One purpose was to determine whether the obtained factor structure of the data resembled the factor structure typically obtained for real test data. The second purpose was to determine whether the obtained unidimensional item characteristics (difficulty and discrimination) were similar to those obtained for real data. The third purpose was to aid in the interpretation of the parameters of the model.

If it were found that a model could not be used to generate realistic data, an attempt was made to determine what changes in the model would yield a more acceptable model. In many cases it was necessary to generate additional data, using different values for the parameters of the model in order to answer specific questions about a particular model statement. Once an understanding was gained as to the roles played by different parameters of the model, predictions could be made regarding the effects of adding or eliminating other terms.

Results. As a result of the analyses performed on the different formulations of the model, a good understanding of the significance of the terms in the model was gained. It is now clear that parameters play quite varied roles depending on the term of the model in which they appear. Because of this, the characteristics of the data for which the model can be used vary markedly, depending on the form of the model.

To begin with, it is clear that the use of $\underset{\sim}{W}'\underset{\sim}{\theta}_j$ and $\underset{\sim}{U}'\underset{\sim}{\sigma}_i$ terms alone is not sufficient for modeling multidimensional response data. The linear composite represented by the $\underset{\sim}{U}'\underset{\sim}{\sigma}_i$ term in the model determines only item difficulty. Moreover, the order of the σ_i vector is unimportant. It is the magnitude of the inner product of the item parameter vector and the weight vector that determines the difficulty of the item. Regardless of whether the vectors have one or five elements, as long as the inner product is the same, the difficulty of the item in terms of proportion of correct responses is the same.

It is also clear from the results of the analyses that the product term $\underset{\sim}{\theta}_j'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ is necessary if item discrimination is to be modeled. When data are generated using only the inner product terms, the items modeled have constant discriminations and the resulting data are unidimensional. When the product term is included, the items modeled have varying discrimination. Moreover, the factor analysis results indicate that the dimensionality of the generated data is determined by the number of elements from the σ_i vector used in the $\underset{\sim}{\theta}_i'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ term. However, it should be emphasized that if the $\underset{\sim}{V}$ matrix contains more than one nonzero element in a row or column, a σ or θ term will appear in the exponent multiplied by more than one of the θ or σ parameters, respectively (e.g., $\theta_1\sigma_1 + \theta_1\sigma_2$ or $\theta_1\sigma_1 + \theta_2\sigma_1$). The presence of these terms in the exponent results in difficulty in determining the meaning of the θ and σ vectors.

The elements in the σ_i vector in the $\underset{\sim}{\theta}_i'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ term determine the discrimination of the modeled items. Because of this, use of the same elements of the σ_i

vector in both the $\underset{\sim}{\theta}_i'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ term and the $\underset{\sim}{U}'\underset{\sim}{\sigma}_i$ term produces an undesirable characteristic in the data. Since the elements used in the $\underset{\sim}{\theta}_i'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ term determine item discrimination, while the elements in the $\underset{\sim}{U}'\underset{\sim}{\sigma}_i$ term determine item difficulty, use of the same elements in both terms yields items having highly related observed item difficulty values and item discrimination values. This is not a very realistic situation.

Conclusions. On the basis of the results of the analyses, it was concluded that if the model is to be used to represent multidimensional item response data, it must include the $\underset{\sim}{\theta}_i'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ term, but no element in either the σ_i or the θ_j vector should be multiplied by more than one term in the other vector. If items are to vary in difficulty, the $\underset{\sim}{U}'\underset{\sim}{\sigma}_i$ term must be included; but to avoid highly related values for unidimensional measures of item discrimination and difficulty, no element of the σ_i vector should appear in both the $\underset{\sim}{U}'\underset{\sim}{\sigma}_i$ term and the $\underset{\sim}{\theta}_i'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ term. The model that appeared to be most useful for modeling multidimensional response data is given by

$$P(x|\underset{\sim}{\theta}_j, \underset{\sim}{\sigma}_i) = \frac{1}{\gamma(\underset{\sim}{\theta}_j, \underset{\sim}{\sigma}_i)} \exp(\underset{\sim}{U}'\underset{\sim}{\sigma}_i + \underset{\sim}{\theta}_j'\underset{\sim}{V}\underset{\sim}{\sigma}_i) , \qquad [21]$$

where no elements of the σ_i vector appear in both terms of the model.

There was one additional significant finding. Although it was concluded that the use of the model without the $\underset{\sim}{\theta}_j'\underset{\sim}{V}\underset{\sim}{\sigma}_i$ term was unsuccessful in modeling multidimensional response data, this result was obtained when the special case of the model was applied to dichotomously scored item response data. This model may also be applied to polychotomously scored item response data. In one specific application of this model to polychotomous response data, some measure of success was attained in modeling multidimensional data using only the $\underset{\sim}{U}'\underset{\sim}{\sigma}_i$ and $\underset{\sim}{W}'\underset{\sim}{\theta}_j$ terms. Dichotomously scored items were transformed to polychotomous form by grouping items together to form clusters having several nominal response categories. When these data were analyzed, several dimensions could be determined. However, this approach has not been extensively investigated, and any conclusions drawn as to the usefulness of this approach are at best tentative.

Estimation of Parameters in the General Rasch Model

Two basic approaches to estimating item parameters can be distinguished. One approach is to specify a distribution of the latent ability of the population from which the sample was taken and then to integrate the response function with respect to that distribution to obtain item parameters unconditionally (Bock, 1972). This approach has been taken by Bock and Lieberman (1970) and Bock and Aitkin (1981). The other approach is to estimate item parameters by treating the examinees' abilities as fixed unknowns and by conditioning item parameter estimation on estimates of ability (Bock, 1972). This approach has been taken by Lord (1968) and Kolakowski and Bock (1970). The present research considers both approaches. However, at this time only the conditional item parameter estimation procedure is designed for the most general form of the model; it is still limited to the use of dichotomously scored data and it does not estimate the $\underset{\sim}{W}$, $\underset{\sim}{U}$, and $\underset{\sim}{V}$ parameter weights. The unconditional item parameter estimation procedure is designed for use with the form of the model given by Equation 21. As was the case with the conditional item parameter estimation procedure, this procedure does not estimate the parameter weights.

In addition to the two procedures that have been developed for item parameter estimation, a maximum likelihood ability estimation procedure has been developed for estimating the ability parameters for the general Rasch model. This procedure estimates ability conditionally and is combined with the conditional maximum likelihood item parameter estimation procedure discussed above to form a conditional maximum likelihood estimation procedure for simultaneous estimation of the item and ability parameters of the general Rasch model. Used alone, the conditional maximum likelihood ability estimation procedure can be used to estimate ability using the item parameter estimates obtained from the unconditional item parameter estimation procedure.

Unconditional Item Parameter Estimation

General procedure. The unconditional item parameter estimation procedure is an adaptation of a procedure proposed by Bock and Aitkin (1981), which was designed for use with a multidimensional 2-parameter normal ogive model (see Equation 13). In the initial step of this procedure, a distribution of ability is assumed, and quadrature nodes and weights are selected for use in determining expected sample sizes for portions of the distribution using numerical integration. For the multidimensional case, the prior distribution of ability is multivariate, and the nodes and weights are vectors. At each node the expected number of examinees from the sample having the ability represented by the node is computed, as is the expected frequency of correct responses to each item by examinees with the ability represented by the node. These expected number-correct scores and expected sample sizes are used in a logit analysis, which is performed using a least squares regression procedure. The results of the logit analysis are estimates of the parameters of the model.

The initial stage of the estimation procedure requires provisional estimates of the item parameters. These provisional estimates are used in the first step of the initial stage, which involves obtaining expected sample sizes and number-correct scores. In the second step of the initial stage, a logit analysis is performed to obtain new estimates of the item parameters. These new estimates are used in the first step of the second stage, which involves obtaining new estimates of sample sizes and number-correct scores. These new sample sizes and number-correct scores are used in another logit analysis, which yields a new set of item parameter estimates. These stages are repeated until a criterion of convergence is met or until a limit on the number of stages is reached.

Expected sample sizes and number-correct scores. The expected sample size at each node is given by

$$\bar{N}_k = \sum_{\ell=1}^{N} \frac{L_\ell(\underset{\sim}{\theta}_k)\underset{\sim}{W}_k}{\tilde{P}_\ell} , \qquad [22]$$

where

$\bar{N}_k$ is the expected sample size at node k;

$\underset{\sim}{\theta}_k$ is the ability represented by node k;

$\underset{\sim}{W}_k$ is the weight for node k;

N is the number of examinees in the sample; and

$L_\ell(\underset{\sim}{\theta}_k)$ is the likelihood of response vector ℓ given node k.

$L_\ell(\underset{\sim}{\theta}_k)$ is given by

$$L_\ell(\underset{\sim}{\theta}_k) = \prod_{i=1}^{n} P(u_{ij}) \ , \qquad [23]$$

where $P(u_{i\ell})$ is given by

$$P(u_{i1}) = \exp[u_{i\ell}(c_i + \underset{\sim}{A}_i\underset{\sim}{\theta}_k)] \ / \ (1 + \exp[c_i + \underset{\sim}{A}_i\underset{\sim}{\theta}_k]) \ , \qquad [24]$$

where

$u_{i\ell}$ is the response to item i in response vector ℓ,

c_i is the difficulty parameter for item i, and

$\underset{\sim}{A}_i\underset{\sim}{\theta}_k$ is given by

$$\underset{\sim}{A}_i\underset{\sim}{\theta}_k = \sum_{m=1}^{M} a_{im}\theta_{km} \ , \qquad [25]$$

where

M is the number of dimensions,

a_{im} is the mth element of the discrimination parameter vector for item i, and

θ_{km} is the mth element of node k.

$\tilde{P}_\ell$ is given by

$$\tilde{P}_\ell = \sum_{k=1}^{q} L_\ell(\underset{\sim}{\theta}_k)\underset{\sim}{W}_k \ , \qquad [26]$$

where q is the number of quadrature nodes. The sum over all of the nodes of the ratio in Equation 22 is one, and the sum of the $\bar{N}_k$ over all of the nodes is N, where N is the number of examinees in the sample.

The expected number-correct score for item i at node k, r_{ik}, is given by

$$\bar{r}_{ik} = \sum_{\ell=1}^{s} \frac{u_{i\ell}L_\ell(\underset{\sim}{\theta}_k)\underset{\sim}{A}_k)}{\tilde{P}_\ell} \ , \qquad [27]$$

where the other terms are as previously defined. The sum over all of the nodes of the r_{ik} for an item is equal to the observed number-correct score for the item.

Multiple logit analysis. For each item the expected proportion-correct score at each node is given by

$$\hat{P}_{ik} = \bar{r}_{ik} / \bar{N}_k , \qquad [28]$$

where all of the terms are as previously defined. The $\hat{P}_{ik}$ are converted to logits and used as the dependent variable in a regression analysis. The independent variables in the regression analysis are the elements in the node vectors. The model for the regression analysis is given by

$$\log_e[\hat{P}_{ik} / (1 - \hat{P}_{ik})] = c_i + \underset{\sim}{A}_i \underset{\sim}{\theta}_k + \text{error} , \qquad [29]$$

where all of the terms are as previously defined. The regression analysis results in estimates of c_i and $\underset{\sim}{A}_i$.

Conditional Item Parameter Estimation

This procedure is based on the conditional maximum likelihood estimation technique. The procedure begins with the computation of an initial weighted item score on each of M dimensions using

$$X_{ik} = \sum_{j=1}^{N} u_{jk} + \sum_{j=1}^{N} \theta_{jk} v_{jk} , \qquad [30]$$

where

X_{ik} is the initial score for item i on dimension k,

u_{jk} is the kth element of $\underset{\sim}{U}$, and

v_{jk} is the kth element of $\underset{\sim}{V}$.

The index j indicates that the value of the elements of $\underset{\sim}{U}$ and $\underset{\sim}{V}$ are dependent on the response of the jth examinee to item i. These scores are converted to z scores via the transformation

$$z_{ik} = (X_{ik} - \bar{X}_k) / s_k , \qquad [31]$$

where

z_{ik} is the z score for item i on dimension k,

X_{ik} is as defined in Equation 30,

$\bar{X}_k$ is the mean of the weighted scores on dimension k, and

s_k is the standard deviation of the weighted scores on dimension k.

The z scores are used as initial estimates of the item parameters. That is,

$$\hat{\sigma}_{0ik} = z_{ik} , \qquad [32]$$

where $\hat{\sigma}_{0ik}$ is the item parameter estimate of item i on dimension k after 0 iterations.

New item parameter estimates are obtained using the initial item parameter estimates as the starting point in an iterative process. One iteration is complete when a new item parameter estimate is obtained for each item on each dimension. Within a single iteration, new estimates on the first dimension are obtained while holding the estimates on all other dimensions constant at the values obtained on the previous iteration. Estimates on the second dimension are obtained using previous estimates on all other dimensions except the first dimension. For the first dimension the new estimates are used. An iteration is complete when new estimates have been obtained on all dimensions. Iterations continue until a criterion of convergence is met or until a limit on the number of iterations is reached.

On the ℓth iteration, the new estimate on the kth dimension for item i is given by

$$\hat{\sigma}_{\ell ik} = \hat{\sigma}_{(\ell-1)ik} + \left[\frac{\partial}{\partial\sigma_{ik}} \log_e L(\sigma_{ik})\right] \Big/ \left[-\frac{\partial^2}{\partial\sigma_{ik}^2} \log_e L(\sigma_{ik})\right], \qquad [33]$$

where $\hat{\sigma}_{(\ell-1)ik}$ is the estimate for item i on dimension k from the previous iteration and log L is the log to the base e of the likelihood function for the response vector for item i. The likelihood function is given by

$$L(\sigma_{ik}) = \prod_{j=1}^{N} P(u_{ij}), \qquad [34]$$

where $P(u_{ij})$ is the probability of response u_{ij} by person j to item i with parameter σ_{ik}. $P(u_{ij})$ is given by

$$P(u_{ij}=1) = \frac{\exp(\underset{\sim}{W}_1'\underset{\sim}{\theta}_j + \underset{\sim}{U}_1'\underset{\sim}{\sigma}_i + \underset{\sim}{\theta}_j'\underset{\sim}{V}_1\underset{\sim}{\sigma}_i + z_1)}{\exp[\underset{\sim}{W}_o'\underset{\sim}{\theta}_j + \underset{\sim}{U}_o'\underset{\sim}{\sigma}_i + \underset{\sim}{\theta}_j'\underset{\sim}{V}_o\sigma_i + z_o] + \exp[\underset{\sim}{W}_1'\underset{\sim}{\theta}_j + \underset{\sim}{U}_1'\underset{\sim}{\sigma}_i + \underset{\sim}{\theta}_j'\underset{\sim}{V}_1\underset{\sim}{\sigma}_i + z_1]}, \qquad [35]$$

where all of the terms are as previously defined. The 0 and 1 subscripts on the vectors of weights, $\underset{\sim}{U}$, $\underset{\sim}{V}$, and $\underset{\sim}{W}$, indicate the values taken by those vectors for an incorrect and a correct response, respectively.

The first derivative of the $\log_e$ likelihood function is given by

$$\frac{\partial}{\partial\sigma_i} \log_e L(\underset{\sim}{\sigma}_i) = \sum_{j=1}^{N} \underset{\sim}{U}_j + \sum_{j=1}^{N} \underset{\sim}{\theta}_j'\underset{\sim}{V}_j - \sum_{j=1}^{N} (\underset{\sim}{U}_o + \underset{\sim}{\theta}_j'\underset{\sim}{V}_o)Q_{ij} - \sum_{j=1}^{N} (\underset{\sim}{U}_1 + \underset{\sim}{\theta}_j'\underset{\sim}{V}_1)P_{ij}, \qquad [36]$$

where all of the items are as previously defined. The second derivative of the $\log_e$ likelihood function is given by

$$\frac{\partial^2}{\partial\sigma_i^2} \log_e L(\underset{\sim}{\sigma}_i) = \sum_{j=1}^{N} [\underset{\sim}{U}_o + \underset{\sim}{U}_1 + \underset{\sim}{\theta}_j'\underset{\sim}{V}_o + \underset{\sim}{V}_o + \underset{\sim}{V}_1]^2 P_{ij}Q_{ij}, \qquad [37]$$

where P_{ij} is the probability of a correct response to item *i* by person *j*, $Q_{ij} = 1 - P_{ij}$, and the other terms are as previously defined.

Conditional Ability Parameter Estimation Procedure

The conditional ability estimation procedure is also a maximum likelihood estimation procedure. It is very similar to the conditional item parameter estimation procedure. For each examinee an initial weighted score is computed on each of M dimensions as

$$X_{jk} = \sum_{i=1}^{n} w_{jk} + \sum_{i=1}^{n} v_{jk}\sigma_{ik} \; , \qquad [38]$$

where X_{jk} is the initial score for person *j* on dimension *k*. These scores are converted to *z* scores via the transformation

$$z_{jk} = (X_{jk} - \bar{X}_k) \; / \; s_k \; , \qquad [39]$$

where

z_{jk} is the *z* score for person *j* on dimension *k*,
X_{jk} is as defined in Equation 38,
$\bar{X}_k$ is the mean of the weighted scores on dimension *k*, and
s_k is the standard deviation of the weighted scores on dimension *k*.

The *z* scores are used as initial estimates of ability. That is,

$$\hat{\theta}_{ojk} = z_{jk} \; , \qquad [40]$$

where $\hat{\theta}_{ojk}$ is the ability estimate of person *j* on dimension *k* after 0 iterations.

Estimates of ability are obtained using the same iterative process as was described for the conditional item parameter estimation procedure. On the ℓth iteration, the new estimate on the *k*th dimension for person *j* is given by

$$\hat{\theta}_{\ell jk} = \hat{\theta}_{(\ell-1)jk} + \left[\frac{\partial}{\partial\theta_{jk}} \log_e L(\theta_{jk})\right] \Bigg/ \left[-\frac{\partial^2}{\partial\theta_{jk}^2} \log_e L(\theta_{jk})\right] , \qquad [41]$$

where $\theta_{(\ell-1)jk}$ is the estimate for person *j* on dimension *k* from the previous iteration and log L is the log to the base *e* of the likelihood function for the response vector for person *j*. The likelihood function is given by

$$L(\theta_{ik}) = \prod_{i=1}^{n} P(u_{ij}) \; , \qquad [42]$$

BEST COPY AVAILABLE

where $P(u_{ij})$ is the probability of a response u_{ij} to item i by person j with ability θ_{jk}. $P(u_{ij})$ is given by Equation 35. In evaluating $\boldsymbol{\theta}_j$ the most recent item parameter estimates are used.

The first derivative of the log likelihood function is given by

$$\frac{\partial}{\partial \boldsymbol{\theta}_j} \log_e L(\boldsymbol{\theta}_j) = \sum_{i=1}^{n} \mathbf{W}_i + \sum_{i=1}^{n} \mathbf{V}_i \boldsymbol{\sigma}_i - \sum_{i=1}^{n} (\mathbf{W}_o + \mathbf{V}_o \boldsymbol{\sigma}_i) Q_{ij} - \sum_{i=1}^{n} (\mathbf{W}_1 + \mathbf{V}_1 \boldsymbol{\sigma}_i) P_{ij} , \qquad [43]$$

where all the terms are as previously defined. The second derivative is given by

$$\frac{\partial^2}{\partial \boldsymbol{\theta}_j^2} \log_e L(\boldsymbol{\theta}_j) = \sum_{i=1}^{n} \left[\mathbf{W}_o + \mathbf{W}_1 + \left(\mathbf{V}_o + \mathbf{V}_1 \right) \boldsymbol{\sigma}_i \right]^2 P_{ij} Q_{ij} , \qquad [44]$$

where, again, all the terms are as previously defined.

Evaluaton of the Estimation Procedures

General Design

The general approach taken to evaluate the estimation procedures was to apply the procedures to test data generated to fit a two-dimensional version of the model given by Equation 24 and then to compare the estimates of the parameters with the values used to generate the simulation data. For this purpose a data set comprising response data for 50 items and 1,000 examinees was generated. Three parameters for each item and two parameters for each examinee were used to generate these data. The values used for the item parameters are shown in Table 1. The examinee ability parameters were selected from a bivariate normal distribution with $\rho = 0$, $\mu = 0$, and Σ equal to the identity matrix.

The weight vectors used in this study were as follows. For an incorrect response, all of the weight vectors were set equal to zero. For a correct response, the following matrix and vectors were used:

$$\mathbf{W} = (0,0) \qquad [45]$$

$$\mathbf{U} = (1,0,0) \qquad [46]$$

$$\mathbf{V} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} . \qquad [47]$$

As can be seen, only the first item parameter was selected by the weight vector, $\mathbf{U}$, to act as the item difficulty parameter. The other two item parameters were selected by the weight matrix, $\mathbf{V}$, to act as discrimination parameters. The resulting model is given by

Table 1
True and Estimated Parameters

Item	True Parameters			Estimates					
				Unconditional			Conditional		
	1	2	3	1	2	3	1	2	3
1	-.65	1.50	.50	-.51	1.45	.44	-.48	1.27	.36
2	-1.40	.50	1.25	-1.34	.48	1.29	-1.56	.45	1.29
3	-.20	1.35	.15	-.02	1.49	.40	-.06	1.69	.25
4	.40	1.60	.55	.64	1.67	.84	.59	1.58	.63
5	.00	.50	1.15	.19	.48	1.02	.10	.51	1.04
6	-1.30	.35	1.05	-1.31	.50	1.17	-1.68	.42	1.25
7	.05	1.45	.35	.28	1.56	.43	.34	1.63	.35
8	.19	.25	1.40	.20	.27	1.18	.25	.20	1.23
9	-.17	.85	.85	-.02	1.07	.87	-.07	.89	.67
10	.14	1.75	.45	.40	1.60	.60	.34	1.45	.57
11	.37	.60	.80	.40	.66	1.02	.29	.61	.88
12	.87	1.65	.65	.82	1.49	.63	.70	1.38	.46
13	-.93	.35	1.35	-.98	.26	1.58	-1.00	.24	1.65
14	1.85	.65	1.65	1.80	.70	1.55	1.59	.54	1.29
15	.06	.65	.65	.16	.73	.66	.00	.80	.52
16	-.41	.45	1.45	-.31	.48	1.53	-.38	.23	1.60
17	-1.54	.75	1.25	-1.41	.72	1.32	-1.55	.71	1.14
18	.34	1.55	.25	.44	1.50	.32	.39	1.46	.29
19	-.15	.65	1.35	.14	.90	1.38	.04	.65	1.28
20	1.48	1.25	.45	1.54	1.41	.41	1.42	1.32	.22
21	-1.45	1.65	.45	-1.42	1.69	.46	-1.73	1.80	.48
22	.75	.45	1.35	.75	.45	1.42	.67	.40	1.35
23	-.75	.35	1.55	-.62	.33	1.75	-.63	.21	1.72
24	1.10	1.10	.30	.99	.92	.37	.93	1.11	.35
25	-.55	1.20	.15	-.42	1.35	.29	-.38	1.37	.24
26	.50	.50	1.00	.44	.66	1.03	.28	.49	1.12
27	-.15	1.45	.45	.10	1.36	.52	-.02	1.32	.42
28	.65	.70	.70	.78	.63	.66	.56	.68	.86
29	-1.00	1.00	.30	-1.02	1.16	.37	-.95	1.14	.38
30	1.00	.30	1.00	1.18	.27	1.27	.98	.27	1.14
31	-.25	.95	.25	-.16	.99	.20	-.04	1.11	.24
32	-.70	.15	1.50	-.65	.17	1.46	-.68	.03	1.51
33	.85	1.15	.45	1.10	1.31	.42	.98	1.29	.30
34	.05	.10	.95	-.02	.13	.92	.03	.20	1.09
35	-.95	1.35	.50	-.79	1.39	.64	-.80	1.28	.52
36	-1.50	.20	1.20	-1.45	.36	1.07	-1.38	.47	1.03
37	1.80	1.55	.55	2.16	1.84	.35	2.06	1.51	.44
38	-2.00	.15	1.15	-1.98	.32	1.05	-2.06	.30	1.09
39	-.90	1.40	.35	-.70	1.38	.22	-.70	1.18	.40
40	1.00	1.00	1.00	1.06	.92	.96	1.04	.87	.79
41	.15	1.25	.70	.26	1.33	.80	.28	1.17	.75
42	-1.50	.25	.95	-1.45	.40	.85	-1.60	.49	1.14
43	-1.25	.35	1.45	-1.02	.27	1.42	-1.13	.26	1.30
44	1.25	1.30	.25	1.37	1.25	.19	.95	1.26	.26
45	-2.00	1.15	.15	-1.89	.99	.09	-2.01	1.30	.13
46	1.75	.50	.50	1.80	.52	.59	1.65	.68	.64
47	.65	.65	1.30	.82	.48	1.39	.71	.49	1.27
48	-.25	1.00	.45	-.19	1.03	.59	-.17	1.03	.53
49	.35	.55	1.15	.40	.59	1.14	.24	.48	1.04
50	.00	.95	.15	-.05	.87	.25	-.07	1.11	.26

$$P(x|\underset{\sim}{\theta}_j,\underset{\sim}{\sigma}_i) = \frac{\exp[\sigma_{i1} + \sigma_{i2}\theta_{j1} + \sigma_{i3}\theta_{j2})x_{ij}]}{1 + \exp[\sigma_{i1} + \sigma_{i1}\theta_{j1} + \sigma_{i3}\theta_{j2}]} . \qquad [48]$$

Results

Table 1 shows the item parameter estimates obtained from both the conditional and unconditional item parameter estimation procedures. The estimates have been scaled to have the same means and standard deviations as the corresponding true item parameters. The correlations of the estimates with the true values are shown in Table 2. As can be seen, for these data there was very little difference in the quality of the estimates yielded by the two estimation procedures. Of course, this comparison is based on simulation data and on only one data set. Clearly, more research is needed before any definite conclusions about these procedures can be drawn.

Table 2
Intercorrelation Matrix for True and Estimated Item Parameters

				Estimation Procedure					
	True			Unconditional			Conditional		
Parameter	σ_{T1}	σ_{T2}	σ_{T3}	$\hat{\sigma}_{U1}$	$\hat{\sigma}_{U2}$	$\hat{\sigma}_{U3}$	$\hat{\sigma}_{C1}$	$\hat{\sigma}_{C2}$	$\hat{\sigma}_{C3}$
σ_{T1}	1.00	.21	-.12	.99	.20	-.12	.99	.15	-.18
σ_{T2}	.21	1.00	-.75	.20	.97	-.72	.15	.96	-.79
σ_{T3}	-.12	-.75	1.00	-.12	-.72	.97	-.18	-.79	.96
$\hat{\sigma}_{U1}$	.99	.20	-.12	1.00	.24	-.13	.99	.19	-.20
$\hat{\sigma}_{U2}$	.20	.97	-.72	.24	1.00	-.73	.19	.96	-.80
$\hat{\sigma}_{U3}$	-.12	-.72	.97	-.13	-.73	1.00	-.20	-.80	.97
$\hat{\sigma}_{C1}$	.99	.15	-.18	.99	.19	-.20	1.00	.20	-.22
$\hat{\sigma}_{C2}$	.15	.96	-.79	.19	.96	-.80	.20	1.00	-.88
$\hat{\sigma}_{C3}$	-.18	-.79	-.96	-.20	-.80	.97	-.22	-.88	1.00

Discussion

The purposes of this paper were threefold. First, the fundamental concepts required when considering multidimensional models for the interaction of a person and a test item were defined. These concepts included the multidimensional latent space, the item difficulty function, and the item discrimination function. These definitions were conceived as multidimensional generalizations of similar concepts in unidimensional IRT models. Second, six existing multidimensional models were reviewed and, on the basis of their similarities, were classified into three general categories. The characteristics of these categories were described, and the general Rasch model was selected for further study on the basis of ease of parameter estimation. Third, estimation procedures for the parameters of the general Rasch model were described and applied to a set of simulation data that had been generated according to a two-dimensional special case of the model. The results indicated that a very close correspondence had been obtained between the estimated item parameters and those used to generate the simulation data.

On the basis of this information, two conclusions can be drawn. First, the concepts of difficulty and discrimination can be generalized to the multidimensional case, but the results are slightly different for compensatory and noncompensatory models. For the compensatory models, the item difficulty is defined by a linear function of the ability dimensions, while for the noncompensatory models, difficulty is defined by a vector of difficulty parameters. In both cases the slope of the item response surface at the difficulty function is a function of the discrimination parameters of the model.

A second conclusion that can be drawn is that the parameters of the general Rasch model can be estimated with acceptable accuracy for the simple two-dimensional case presented in the paper. This is, of course, very minimal evidence for the value of the estimation procedures; but combined with the work of Bock and Aitkin (1981), the results look fairly promising.

The results presented in this paper summarize only the initial steps in a thorough study of the applicability and usefulness of multidimensional latent trait models. Much further work needs to be done. The sample size requirements and the characteristics of the estimates obtained from the estimation procedures need to be determined. Procedures for determining the number of dimensions required for the multidimensional latent space must be developed as well as guidelines for interpreting the dimensions. The usefulness of the models for real data applications such as test construction and adaptive testing should be investigated. Also the advantages of these procedures over existing multidimensional procedures, such as factor analysis, should be studied.

REFERENCES

American College Testing Program. Content of the tests in the ACT Assessment. Iowa City IA: Author, 1980.

Andersen, E. B. A general latent structure model for contingency table data (Research Report No. 82). Copenhagen: Universitetets Statistiske Institut, 1982.

Andrich, D. A rating formulation for ordered response categories. Psychometrika, 1978, 43, 561-573.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Bock, R. D., & Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 1981, *46*, 443-459.

Bock, R. D., & Lieberman, M. Fitting a response model for *n* dichotomously scored items. *Psychometrika*, 1970, *35*, 179-197.

Christoffersson, A. Factor analysis of dichotomized variables. *Psychometrika*, 1975, *40*, 5-32.

Holzman, T. G., Glaser, R., & Pellegrino, J. W. *Cognitive determinants of series completion: Individual and developmental differences.* Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Kim, J., & Mueller, C. W. *Factor analysis: Statistical methods and practical issues.* Beverly Hills CA: Sage Publications, 1978.

Kolakowski, D., & Bock, R. D. *A Fortran IV program for maximum likelihood item analysis and test scoring: Normal ogive model* (Research Memo No. 12). Chicago: The University of Chicago, Department of Education, Statistical Laboratory, 1970.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F. M. Discussion: Session 2. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Lord, F. M. *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading MA: Addison-Wesley, 1968.

McKinley, R. L., & Reckase, M. D. *Multidimensional latent trait models.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, March 1982.

Mulaik, S. A. *A mathematical investigation of some multidimensional Rasch models for psychological tests.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ, March 1972.

Muthén, B. Contributions to factor analysis of dichotomous variables. *Psychometrika*, 1978, *43*, 551-560.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danske Paedagogike Institut, 1960.

Rasch, G. On general laws and the meaning of measurement in psychology. In J.

Neyman, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4). Berkeley: University of California Press, 1961.

Reckase, M. D. Development and application of a multivariate logistic latent trait model (Doctoral dissertation, Syracuse University, 1972). *Dissertation Abstracts International*, 1973, *33*. (University Microfilms No. 73-7762)

Reckase, M. D. *Ability estimation and item calibration using the one and three parameter logistic models: A comparative study* (Research Report 77-1). Columbia: University of Missouri, Department of Educational Psychology, November 1977.

Reckase, M. D. *The formation of homogeneous item sets when guessing is a factor in item responses* (Research Report 81-5). Columbia: University of Missouri, Department of Educational Psychology, August 1981.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, 1974, *39*, 111-121.

Sympson, J. B. A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Whitely, S. E. Multicomponent latent trait models for ability tests. *Psychometrika*, 1980, *45*, 479-494.

DISCUSSION

FRITZ DRASGOW
UNIVERSITY OF ILLINOIS

The research presented by Reckase and McKinley, as well as the work of Bock and Aitkin (1981), Christofferson (1975), and Muthén (1978), has led me to think about why multidimensional models for item responses are important. Perhaps the most obvious argument for their importance is based on the general principle that it is desirable to use a model that is "correct" or appropriate for the data when conducting a statistical analysis. Because tests are usually multidimensional, it would follow that multidimensional models are required. However, I think it would be a serious mistake if item response theorists accepted this argument. My reasoning is based on results from the factor analysis literature. Humphreys (1982), for example, noted that "tests of seemingly high homogeneity can frequently be 'splintered' into several different tests" (p. 3) and therefore it is possible to identify "literally thousands of factors" (p. 3).

Is there any reason to measure this vast number of highly redundant latent traits? From the standpoints of substantive theory and applied prediction, it appears that splintering factors is useful only when the splintered factors show differential patterns of relations with other important variables. By this rule for parsimony, only a very small number of latent traits seems to be required (see Humphreys, 1982). These latent traits are broad attributes of individuals such as verbal reasoning and quantitative reasoning. One advantage of restricting attention to a few broadly based latent traits is that usually it is not too difficult to determine what trait a particular item measures from a consideration of item content.

I do think that multidimensional models are essential in situations where two or more distinctly different latent traits influence item responses. Consider, for example, a test of quantitative reasoning administered to a sample that includes individuals for whom English is a second language. It would be very interesting to obtain estimates of quantitative ability that were not affected by English fluency. Similarly, it would be interesting to separate mechanical aptitude from proficiency with paper and pencil when examinees are administered a paper-and-pencil test of mechanical aptitude. This application is particularly significant because the United States judiciary appears to be adopting the position that paper-and-pencil tests primarily measure paper-and-pencil aptitude rather than vocational aptitude.

To summarize, multidimensional item response theory models may be able to provide solutions to substantive and applied prediction problems that are very difficult to address with unidimensional item response theory. Item response theorists should be careful, however, to remain aware of the fact that attributes of individuals can usually be splintered into more homogeneous attributes without concomitant gains in the understanding and accuracy of prediction of other behaviors.

References

Bock, R. D., & Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 1981, *46*, 443-459.

Christofferson, A. Factor analysis of dichotomized variables. *Psychometrika*, 1975, *40*, 5-32.

Humphreys, L. G. *Systematic heterogeneity of items in tests of meaningful and important psychological attributes: A rejection of unidimensionality.* Paper presented at the Fifth International Conference on Educational Testing, University of Stirling, 1982.

Muthén, B. Contributions to factor analysis of dichotomous variables. *Psychometrika*, 1978, *43*, 551-560.

ESTIMATION OF ITEM PARAMETERS AND THE GEM ALGORITHM

ROBERT K. TSUTAKAWA
UNIVERSITY OF MISSOURI

Under the assumption that ability parameters are randomly sampled from a prior distribution with unknown parameters, various empirical Bayes type estimators of ability and item parameters were presented by Rigdon and Tsutakawa (1981). Their numerical work using the 1-parameter logistic model suggested that extensions to multiparameter models could be computationally too costly if there is interest in obtaining posterior distributions of both ability and item parameters. Here, concern will be with one of these estimators, namely an MLF procedure (a maximum likelihood procedure where the item parameters are treated as fixed parameters, as opposed to a procedure which assumes they have prior distributions), where extensions to multiparameter models appear straightforward and asymptotic methods are available for studying posterior distributions of item parameters.

The models and procedures discussed in this paper are related to those presented in Bock and Aitkin (1981), where further references may be found. Bock and Aitkin considered the 2-parameter probit model and approximated a normally distributed prior distribution of abilities by a finite and discrete distribution. They successfully used the EM algorithm by treating as unobserved data the number of examinees at each ability level and the number of these examinees who responded correctly to each item. Although they mentioned the possible use of the general EM (GEM) algorithm, they cautioned that the principle is not yet well established due to an error in a proof by Dempster, Laird, and Rubin (1977). Wu (1983) has pointed out that it was an incorrect use of the triangle inequality which invalidates Dempster, Laird, and Rubin's (1977) proof of the convergence of the EM algorithm.

One purpose of this paper is to clarify the nature of the GEM solution, assuming that convergence has already taken place. For this purpose the general situation will first be considered and conditions then given under which the GEM solution maximizes the likelihood function based on incomplete data. For the 2-parameter logistic model, the equations occurring at each iteration of the GEM algorithm will be compared with the likelihood equations for the incomplete data. It will be shown why the GEM approach is computationally simpler than the solution via direct methods. One question which remains is, under what conditions is there convergence? In practice, for latent trait applications in particular, once there is convergence, it is usually quite easy to test the solution by examining the likelihood function in a neighborhood of the solution and to verify whether it is at least a local maximum. The paper concludes by showing that for the 1-parameter logistic model, convergence is assured by the concavity of the log-likelihood function.

Note on the General EM Application

Dempster, Laird, and Rubin (DLR; 1977), have described the GEM algorithm for finding the maximum likelihood estimator based on the incomplete data y when the distribution of the complete data x does not belong to the exponential family. Following their notation, let $f(x|\phi)$ be the pdf of x depending on the r-dimensional parameter ϕ belonging to a convex set Ω in E^r and let $g(y|\phi)$ be the pdf of the incomplete data, i.e.,

$$g(y|\phi) = \int_{X(y)} f(x|\phi)dx \qquad [1]$$

where $X(y)$ denotes the set of x mapping into the point y. Moreover, let

$$Q(\phi'|\phi) = E\{\log f(x|\phi')|y,\phi\} \ , \qquad [2]$$

the conditional expectation of the log likelihood at ϕ' when ϕ obtains and y is given. An instance of the GEM algorithm starts with some value ϕ and iteratively maximizes $Q(\phi'|\phi)$ with respect to ϕ' while updating the value of ϕ by $M(\phi)$, the maximizing value of ϕ' at the end of each iteration, until convergence is attained.

Suppose the algorithm converges to some value ϕ^*, in the sense that the successive values of ϕ' so computed converge to ϕ^*. Does ϕ^* maximize $g(y|\phi)$? Since the algorithm deals with Q and not directly with g, it is not clear how the solution relates to g. This relation is clarified below.

For each $\phi \in \Omega$, let

$$L(\phi) = \log g(y|\phi) \ , \qquad [3]$$

$$k(x|y,\phi) = f(x|\phi)/g(y|\phi) \ , \qquad [4]$$

for $x \in X(y)$, and for each $(\phi',\phi) \in \Omega\times\Omega$, let

$$H(\phi'|\phi) = \int_{X(y)} [\log k(x|y,\phi')]k(x|y,\phi)dx \ . \qquad [5]$$

Then, as shown by DLR, for any $(\phi,\phi^\circ) \in \Omega\times\Omega$

$$L(\phi) = Q(\phi|\phi^\circ) - H(\phi|\phi^\circ) \ . \qquad [6]$$

If the derivatives exist and $k(x|y,\phi)$ is regular in the sense of Wilks (1962), it follows from Equation 6 that

$$\left.\frac{\partial L(\phi)}{\partial \phi_j}\right|_{\phi=\phi^\circ} = \left.\frac{\partial Q(\phi|\phi^\circ)}{\partial \phi_j}\right|_{\phi=\phi^\circ} \qquad [7]$$

since

$$\left.\frac{\partial H(\underset{\sim}{\phi}|\underset{\sim}{\phi}^{\circ})}{\partial \phi_j}\right|_{\underset{\sim}{\phi}=\underset{\sim}{\phi}^{\circ}} = 0 \qquad [8]$$

if $k(\underset{\sim}{x}|\underset{\sim}{y},\underset{\sim}{\phi})$ is regular, $j = 1, \ldots, k$.

Thus, if $L(\underset{\sim}{\phi})$ has a mode $\hat{\underset{\sim}{\phi}}$, which is also the unique solution to the likelihood equations,

$$\frac{\partial L(\underset{\sim}{\phi})}{\partial \phi_j} = 0 \ , \ j = 1, \ldots, \qquad [9]$$

then by Equation 7 the solution $\underset{\sim}{\phi}^{\circ}$ to the equation

$$\left.\frac{\partial Q(\underset{\sim}{\phi}|\underset{\sim}{\phi}^{\circ})}{\partial \phi_j}\right|_{\underset{\sim}{\phi}=\underset{\sim}{\phi}^{\circ}} = 0 \ , \ j = 1, \ldots, r \qquad [10]$$

will also be the mode of $L(\underset{\sim}{\phi})$, i.e., $\hat{\underset{\sim}{\phi}} = \underset{\sim}{\phi}^{\circ}$.

Moreover, if the algorithm converges to $\underset{\sim}{\phi}^*$, then $M(\underset{\sim}{\phi}^*) = \underset{\sim}{\phi}^*$, and

$$Q(M(\underset{\sim}{\phi}^*)|\underset{\sim}{\phi}^*) = Q(\underset{\sim}{\phi}^*|\underset{\sim}{\phi}^*) \ ; \qquad [11]$$

therefore,

$$\left.\frac{\partial Q(\underset{\sim}{\phi}|\underset{\sim}{\phi}^*)}{\partial \phi_j}\right|_{\underset{\sim}{\phi}=\underset{\sim}{\phi}^*} = 0 \ , \ j = 1, \ldots, r. \qquad [12]$$

Thus, $\hat{\underset{\sim}{\phi}} = \underset{\sim}{\phi}^*$ under the above uniqueness condition.

Statistical Model for Item Responses

Consider the n × k matrix $\underset{\sim}{Y}$ of binary responses Y_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, k$ where $Y_{ij} = 0$ or 1 depending whether the ith examinee's response to item j is incorrect or correct. Let

$$p_{ij} = p(Y_{ij}=1|\theta_i, \underset{\sim}{\beta}_j) \qquad [13]$$

be a probability model for responses depending on θ_i, a real valued ability (or latent trait) parameter of the ith examinee, and β_j, a possibly vector-valued item parameter of the jth item. Given $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_n)^T$ and $\underset{\sim}{\beta} = (\beta_1, \ldots, \beta_k)^T$, conditional independence is assumed among the responses in $\underset{\sim}{Y}$ in the sense that the conditional joint distribution of $\underset{\sim}{Y}$, given $(\underset{\sim}{\theta},\underset{\sim}{\beta})$ is

$$p(\underset{\sim}{y}|\underset{\sim}{\theta},\underset{\sim}{\beta}) = \prod_{i=1}^{n}\prod_{j=1}^{k} p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}} \qquad [14]$$

Further assume that $\underset{\sim}{\theta}$ may be treated as a random sample from some prior distribution, $p(\theta|\underset{\sim}{\lambda})$, having parameter $\underset{\sim}{\lambda}$.

Let $\underset{\sim}{Y}_i = (Y_{i1}, \ldots, Y_{ik})$, the response to the $\underline{k}$ items by person $\underline{i}$. Then, the unconditional probability of responses by person $\underline{i}$ is

$$p(\underset{\sim}{y}_i|\underset{\sim}{\beta},\underset{\sim}{\lambda}) = \int p(\theta_i|\underset{\sim}{\lambda}) \prod_{j=1}^{k} p(y_{ij}|\theta_i,\beta_j)d\theta_i \qquad [15]$$

and the joint probability of $\underset{\sim}{Y}$ is

$$p(\underset{\sim}{y}|\underset{\sim}{\beta},\underset{\sim}{\lambda}) = \prod_{i=1}^{n} p(\underset{\sim}{y}_i|\underset{\sim}{\beta},\underset{\sim}{\lambda}) \ . \qquad [16]$$

There are certain uniqueness problems associated with this parameterization. These problems are usually handled by placing constraints on the parameters or by reparameterization. In the interest of keeping the discussion simple, these problems will be ignored except when discussing specific models.

Estimation of Item Parameters

As formulated above, there is a sequence of $\underline{n}$ independent and identically distributed response vectors $\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_n$ with likelihood function

$$\ell(\underset{\sim}{\beta},\underset{\sim}{\lambda}) = \prod_{i=1}^{n} p(\underset{\sim}{y}_i|\underset{\sim}{\beta},\underset{\sim}{\lambda}) \ , \qquad [17]$$

of the parameter $\underset{\sim}{\phi} = (\underset{\sim}{\beta},\underset{\sim}{\lambda})$. Now consider the maximum likelihood estimator $\hat{\underset{\sim}{\phi}}$, which is the value of $\underset{\sim}{\phi}$ that maximizes the likelihood function. In practice, this estimator is not trivially obtained. Even for moderate size $\underline{n}$ and $\underline{k}$, direct maximization by Newton-Raphson type methods is likely to be too costly (cf. Bock & Lieberman, 1970). Since the joint distribution of $(\underset{\sim}{Y},\underset{\sim}{\theta})$ generally does not belong to an exponential family, a direct application of the EM algorithm is generally not feasible. Bock and Aitkin (1981) have successfully avoided this problem by approximating the prior by a known discrete distribution. The implementation of the GEM algorithm, which extends the EM algorithm to nonexponential families, was demonstrated for latent trait models by Rigdon and Tsutakawa (1981). The fact that GEM solutions are indeed maximum likelihood estimators follows under certain conditions from the general properties given in the second section of this paper. These properties will now be discussed further in terms of the 2-parameter logistic model.

Consider the 2-parameter logistic model defined by

$$p(y_{ij}|\theta_i,a_j,b_j) = \frac{e^{y_{ij}(a_j+b_j\theta_i)}}{1 + e^{(a_j+b_j\theta_i)}} \tag{18}$$

where $(\theta_1, \ldots, \theta_n)$ is a random sample from $N(0, 1)$, $b_j > 0$ and the previous $\underset{\sim}{\beta}_j$ is now (a_j, b_j). This model is equivalent to the more familiar model

$$p(y_{ij}|\theta_i',\alpha_j,\beta_j) = \frac{e^{y_{ij}\alpha_j(\theta_i'-\beta_j)}}{1 + e^{\alpha_j(\theta_i'-\beta_j)}} \tag{19}$$

with the constraints $\Sigma_{j=1}^{k}\beta_j = 0$ and $\Pi_{j=1}^{k}\alpha_j = 1$, where $(\theta_1', \ldots, \theta_n')$ is a random sample from $N(\mu,\sigma^2)$ with both μ and σ^2 unknown. The equivalence may be seen by the relation

$$\theta_i = (\theta_i'-\mu)/\sigma \ ,$$

$$a_j = -\alpha_j(\beta_j-\mu) \ , \tag{20}$$

and

$$b_j = \sigma\alpha_j \ .$$

The joint probability of $\underset{\sim}{y}$ is then given by

$$g(\underset{\sim}{y}|\underset{\sim}{\beta}) = \prod_{i=1}^{n} \int \prod_{j=1}^{k} p(y_{ij}|\theta_i,a_j,b_j)p(\theta_i)d\theta_i \ , \tag{21}$$

where $p(\theta_i) = [1/(2\pi)^{1/2}]\exp(-\theta_i^2/2)$ and the posterior pdf of θ_i by

$$p(\theta_i|\underset{\sim}{y},\underset{\sim}{\beta}) = p(\theta_i|\underset{\sim}{y}_i,\underset{\sim}{\beta})$$

$$\propto p(\theta_i) \prod_{j=1}^{k} p(y_{ij}|\theta_i,a_j,b_j) \ . \tag{22}$$

Moreover, the partial derivatives of the log likelihood are

$$\frac{\partial}{\partial a_J} \log g(\underset{\sim}{y}|\underset{\sim}{\beta})$$

$$= \sum_{i=1}^{n} \frac{\int p(\theta_i) \prod_{j=1}^{k} p(y_{ij}|\theta_i,a_j,b_j)\frac{\partial}{\partial a_J} \log p(y_{iJ}|a_J,b_J,\theta_i)d\theta_i}{g(\underset{\sim}{y}_i|\underset{\sim}{\beta})}$$

$$= \sum_{i=1}^{n} \int \left\{ y_{iJ} - \frac{1}{1 + \exp(-a_J - b_J\theta_i)} \right\} p(\theta_i | \underset{\sim}{y}_i, \underset{\sim}{\beta}) d\theta_i$$

$$= \sum_{i=1}^{n} y_{iJ} - \Sigma\, E \left\{ \frac{1}{1 + \exp(-a_J - b_J\theta_i)} \Big| \underset{\sim}{y}_i, \underset{\sim}{\beta} \right\} \qquad [23]$$

and

$$\frac{\partial \log g(\underset{\sim}{y} | \underset{\sim}{\beta})}{\partial b_J}$$

$$= \sum_{i=1}^{n} \frac{\int p(\theta_i) \prod_{j=1}^{k} p(y_{ij} | \theta_i, a_j, b_j) \frac{\partial}{\partial b_J} \log p(y_{iJ} | \theta_i, a_J, b_J) d\theta_i}{g(\underset{\sim}{y}_i | \underset{\sim}{\beta})}$$

$$= \sum_{i=1}^{n} \int \left\{ y_{iJ}\theta_i - \frac{\theta_i}{1 + \exp(-a_J - b_J\theta_i)} \right\} p(\theta_i | \underset{\sim}{y}_i, \underset{\sim}{\beta}) d\theta_i$$

$$= \sum_{i=1}^{n} y_{iJ}\, E\{\theta_i | \underset{\sim}{y}_i, \underset{\sim}{\beta}\} - \sum_{i=1}^{n} E \left\{ \frac{\theta_i}{1 + \exp(-a_J - b_J\theta_i)} \Big| \underset{\sim}{y}_i, \underset{\sim}{\beta} \right\} \qquad [24]$$

where $g(\underset{\sim}{y}_i | \underset{\sim}{\beta})$ is the marginal of $\underset{\sim}{y}_i$, $J = 1, \ldots, k$. Thus, the likelihood equations are

$$q_J = \sum_{i=1}^{n} E \left\{ \frac{1}{1 + \exp(-a_J - b_J\theta_i)} \Big| \underset{\sim}{y}_i, \underset{\sim}{\beta} \right\} \qquad [25]$$

and

$$\sum_{i=1}^{n} E\{y_{iJ}\theta_i | \underset{\sim}{y}_i, \underset{\sim}{\beta}\} = \sum_{i=1}^{n} E \left\{ \frac{\theta_i}{1 + \exp(-a_J - b_J\theta_i)} \Big| \underset{\sim}{y}_i, \underset{\sim}{\beta} \right\}, \qquad [26]$$

$J = 1, \ldots, k$, where $q_J = \Sigma_{i=1}^{n} y_{iJ}$.

On the other hand, consider the maximization of $Q(\underset{\sim}{\beta}' | \underset{\sim}{\beta})$ with respect to $\underset{\sim}{\beta}'$ in a given iteration of the GEM algorithm applied to the 2-parameter logistic model. (Note that now $\underset{\sim}{\phi} = \underset{\sim}{\beta}$ and $\underset{\sim}{\phi}' = \underset{\sim}{\beta}'$, since λ does not exist under the current parameterization.) Now

$$Q(\underset{\sim}{\beta}' | \underset{\sim}{\beta}) = E \left\{ \sum_{i=1}^{n} \log p(\theta_i) | \underset{\sim}{\beta}, \underset{\sim}{y} \right\} + E \left\{ \sum_{ij}\Sigma \log p(y_{ij} | \theta_i, a_j', b_j') | \underset{\sim}{y}, \underset{\sim}{\beta} \right\} . \qquad [27]$$

Setting the partial derivatives of $Q(\underset{\sim}{\beta}'|\underset{\sim}{\beta})$ with respect to a'_J and b'_J equal to zero gives the equations

$$q_J = \sum_{i=1}^{n} E\left\{ \frac{1}{1 + \exp(-a'_J - b'_J\theta_i)} \middle| \underset{\sim}{y}_i, \underset{\sim}{\beta} \right\} \qquad [28]$$

and

$$\sum_{i=1}^{n} E\{y_{iJ}\theta_i | \underset{\sim}{y}_i, \underset{\sim}{\beta}\} = \sum_{i=1}^{n} E\left\{ \frac{\theta_i}{1 + \exp(-a'_J - b'_J\theta_i)} \middle| \underset{\sim}{y}_i, \underset{\sim}{\beta} \right\} \qquad [29]$$

$J = 1, \ldots, k$.

Note the similarities between the likelihood equations and the GEM equations, i.e., Equations 25 and 26 vs. Equations 28 and 29. The important difference between Equations 25 and 28 is that in Equation 28 the conditional expectation is with respect to $\underset{\sim}{y}_i$ and a fixed value of $\underset{\sim}{\beta}$ obtained from the previous iteration; whereas in Equation 25 the conditional expectation is with respect to $\underset{\sim}{y}_i$ and the variable $\underset{\sim}{\beta}$, which also occurs in the expectation, since $\underset{\sim}{\beta} = \{(a_1, b_1), \ldots, (a_k, b_k)\}$. A similar difference can be noted between Equations 26 and 29.

Note also that since $\underset{\sim}{\beta}$ is fixed in Equations 28 and 29, the value for (a'_J, b'_J) can be derived separately for each J, whereas with Equations 25 and 26, (a_J, b_J) must be found simultaneously for all J, since the equations for different J are related through a common $\underset{\sim}{\beta}$. It is thus seen that the simplicity of the GEM approach is due to the considerably simpler equations to be solved.

If there is convergence of the GEM algorithm to some value $\underset{\sim}{\beta}^*$, then $M(\underset{\sim}{\beta}^*) = \underset{\sim}{\beta}^*$. In this case Equations 28 and 29 are satisfied by taking $\underset{\sim}{\beta} = \underset{\sim}{\beta}^*$ and $(a'_1, b'_1), \ldots, (a'_k, b'_k) = \underset{\sim}{\beta}^*$. Moreover, $\underset{\sim}{\beta}^*$ satisfies Equations 25 and 26. Thus, if there is a unique solution to the likelihood equations and this is the mode, then the GEM solution will be this mode.

On the Concavity of the Log-Likelihood Function

Conditions under which the GEM algorithm produces the maximum likelihood estimate $\hat{\underset{\sim}{\phi}} = (\hat{\underset{\sim}{\beta}}, \hat{\underset{\sim}{\lambda}})$ are still unknown. However, if the logarithm of the likelihood function of Equation 3 is concave and the algorithm converges to a finite value, then this value is the maximum likelihood estimate. An important example where such a property holds is the 1-parameter model

$$P_{ij} = 1/[1 + e^{-(\theta_i - \beta_j)}] , \qquad [30]$$

and when $\theta_1, \ldots, \theta_n$ are independently and identically distributed, $N(0,\sigma^2)$.

The joint distribution of the complete data $(\underset{\sim}{\theta},\underset{\sim}{Y})$ is then

$$f(\underset{\sim}{\theta},\underset{\sim}{y}|\underset{\sim}{\beta},h) = \prod_{i=1}^{n} \frac{h^{1/2}}{\sqrt{2\pi}} \exp(-h\theta_i^2/2) \prod_{j=1}^{k} \frac{e^{y_{ij}(\theta_i-\beta_j)}}{1+e^{(\theta_i-\beta_j)}}, \qquad [31]$$

where $h = \sigma^{-2}$. Now, it is well known that

$$\frac{h^{1/2}}{\sqrt{2\pi}} \exp(-h\theta_i^2/2) \qquad [32]$$

is log concave with respect to $\underline{h}$. Moreover, it is shown by Pratt (1981, p. 105) that

$$e^{y_{ij}(\theta_i-\beta_i)} \Big/ [1 + e^{(\theta_i-\beta_j)}] \qquad [33]$$

is log concave with respect to (β_j,θ_i) for $y_{ij} = 0, 1$. Since the product of log concave functions is log concave, it follows that $f(\theta,\underset{\sim}{y}|\underset{\sim}{\beta},h)$ is log concave with respect to (θ,β,h). According to the theorem quoted by Pratt (1981, p. 105, §3), an integral of a log concave function with respect to any of its arguments is log concave with respect to the remaining arguments. This theorem implies that

$$g(\underset{\sim}{y}|\underset{\sim}{\beta},\underset{\sim}{h}) = \int f(\underset{\sim}{\theta},\underset{\sim}{y}|\underset{\sim}{\beta},h)\, d\underset{\sim}{\theta} \qquad [34]$$

is log concave with respect to $(\underset{\sim}{\beta},h)$. Thus, if the GEM algorithm converges to a finite value for the 1-parameter logistic with the normal prior on ability, the resulting solution is indeed the maximum likelihood estimate.

REFERENCES

Bock, R. D., & Aitkin, M. Marginal maximum likelihood estimator of item parameters: An application of an EM algorithm. Psychometrika, 1981, 46, 443-459.

Bock, R. D., & Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). Journal of the Royal Statistical Society, Series B, 1977, 39, 1-38.

Pratt, J. W. Concavity of the log likelihood. Journal of the American Statistical Association, 1981, 76, 103-106.

Rigdon, S. E., & Tsutakawa, R. K. Estimation in latent trait models (Research Report 81-1; Mathematical Sciences Technical Report No. 102). University of Missouri, Department of Statistics, 1981.

Wilks, S. S. Mathematical statistics. New York: John Wiley & Sons, 1962.

Wu, C.-F. On the convergence properties of the EM algorithm. The Annals of Statistics, 1983, 11, 95-103.

ACKNOWLEDGMENTS

This study was prepared under Contract No. N00014-81-K-0265, NR 150-464, with the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research.

IMPLEMENTATION OF THE EM ALGORITHM IN THE ESTIMATION OF ITEM PARAMETERS: THE BILOG COMPUTER PROGRAM

ROBERT J. MISLEVY
INTERNATIONAL EDUCATIONAL SERVICES

R. DARRELL BOCK
UNIVERSITY OF CHICAGO

Marginal maximum likelihood equations for estimating the item parameters in the 1- and 2-parameter normal ogive item response models were introduced by Bock and Aitkin (1981). The iterative solution of these equations bears strong resemblence to the EM algorithm of Dempster, Laird, and Rubin (1977). Over the past year, similar procedures have been implemented in the BILOG computer program (Bock & Mislevy, 1982) for estimating item parameters in the 1-, 2-, and 3-parameter logistic ogive models. Extensions of the original Bock and Aitkin solution include the simultaneous characterization of the latent population distribution and the incorporation of Bayes priors on item parameters, so that Bayes modal rather than maximum likelihood estimates may be obtained.

The purpose of this paper is to review the basic elements of the EM approach to estimating item parameters and to illustrate its use with one simulated and one real data set. The examples bring into focus a topic of occasional discussion in psychometric circles, namely, the degree to which item parameters in the 3-parameter model can be recovered.

An EM Algorithm for Estimating Item Parameters

The 3-parameter logistic ogive item response model for dichotomous test items, of which the 1- and 2-parameter models may be considered special cases, expresses the probability that person $\underline{i}$ will respond correctly to item $\underline{j}$ as

$$\begin{aligned} P_{ij} &= \mathrm{Prob}(x_{ij}=1) \\ &= G_j + (1-G_j)\,\Psi[A_j(\theta_i-B_j)] \\ &= G_j + (1-G_j)\,\Psi[A_j\theta_i + C_j]\ , \end{aligned} \qquad [1]$$

where

x_{ij}, the item response, is 1 if correct and 0 if incorrect;
$\Psi(x)$ is the cumulative logistic function; $1/[1 + \exp(-x)]$;

G_j is the lower asymptote, often called the guessing parameter of item j, identically zero in the 1- and 2-parameter models;

A_j is the slope of item j, a constant over items in the 1-parameter model;

B_j is the threshold of item j;

C_j, equal to $-A_jB_j$, is the item intercept, introduced because estimation equations for the intercepts are simpler than those for item thresholds; and

θ_i is the ability of person i.

Given observed responses x_{ij} from N persons to n items, item parameters may be estimated. The main problem arising in this endeavor is that except in the 1-parameter model, the person parameters cannot be eliminated from the maximum likelihood estimation equations of the item parameters. In the presence of the so-called "nuisance" parameters, the standard results of maximum likelihood theory (e.g., consistency) do not apply.

A Solution When Ability Is Known

Estimation of item parameters would be straightforward if person ability values were known rather than implied by item responses. This is essentially the case that obtains in the bioassay setting, where the researcher controls the level of treatment dosage to each experimental unit, observes the proportion of units exhibiting the targeted response at each dosage level, then estimates an hypothesized underlying logistic or normal response function. In anticipation of the EM solution for item parameters, likelihood equations are presented for a logit regression problem that parallels the psychometric problem.

Suppose that, as in the bioassay setting, responses to each of n test items are observed from groups of persons at each of q specified points along the ability scale. Let N_{jk} be the number of responses to item j from persons with ability X_k and let R_{jk} be the number of these responses that are correct. Under the usual assumption of local independence, the total likelihood of a collection of observations of this type is as follows:

$$L = \prod_j \prod_k \frac{N_{jk}!}{(N_{jk}-R_{jk})!\ R_{jk}!} P_{jk}^{R_{jk}} (1-P_{jk})^{N_{jk}-R_{jk}} \qquad [2]$$

where

$$P_{jk} = G_j + (1-G_j)\ \Psi(A_jX_k + C_j)\ . \qquad [3]$$

The likelihood equations for the item parameters are the first derivatives of the log of Equation 2, equated to zero:

$$C_j: \quad 0 = \sum_k (R_{jk} - P_{jk}N_{jk})W_{jk} \qquad [4]$$

$$A_j: \quad 0 = \sum_k (R_{jk} - P_{jk}N_{jk})W_{jk}X_{jk} \qquad [5]$$

$$G_j: \quad 0 = (1-G_j)^{-1} \sum_k (R_{jk} - P_{jk}N_{jk})/P_{jk} \qquad [6]$$

where

$$W_{jk} = \frac{(1-G_j)P^*_{jk}(1-P^*_{jk})}{P_{jk}(1-P_{jk})} \qquad [7]$$

with

$$P_{jk}{}^* = \Psi(A_jX_k + C_j) \ . \qquad [8]$$

If the vector of zeros that solves these equations is unique and if the matrix of second derivatives of the log of Equation 2 is positive definite when evaluated at these values, then these values are the maximum likelihood estimates of the item parameters. The second derivatives are

$$C_j,C_j: \quad \sum_k P^*_{jk}(1-P^*_{jk})(G_jR_{jk}/P^2_{jk} - N_{jk}) \qquad [9]$$

$$C_j,A_j: \quad \sum_k P^*_{jk}(1-P^*_{jk})(G_jR_{jk}/P^2_{jk} - N_{jk})X_{jk} \qquad [10]$$

$$C_j,G_j: -\sum_k P^*_{jk}(1-P^*_{jk})R_{jk}/P^2_{jk} \qquad [11]$$

$$A_j,A_j: \quad \sum_k P^*_{jk}(1-P^*_{jk})(G_jR_{jk}/P^2_{jk} - N_{jk})X^2_{jk} \qquad [12]$$

$$A_j,G_j: -\sum_k P^*_{jk}(1-P^*_{jk})R_{jk}X_{jk}/P^2_{jk} \qquad [13]$$

$$G_j,G_j: \quad (1-G_j)^{-2} \sum_k [R_{jk}/P_{jk} - N_{jk} - R_{jk}(1-P_{jk})/P^2_{jk}] \qquad [14]$$

The solution of the likelihood equations may be accomplished by Newton-Raphson iterations, carried out item by item. The t + 1th iteration is

$$\begin{bmatrix} \hat{C}_j{}^{t+1} \\ \hat{A}_j{}^{t+1} \\ \hat{G}_j{}^{t+1} \end{bmatrix} = \begin{bmatrix} \hat{C}_j{}^{t} \\ \hat{A}_j{}^{t} \\ \hat{G}_j{}^{t} \end{bmatrix} - \begin{bmatrix} \mathrm{SDRV}(C_jC_j) & \mathrm{SDRV}(A_jC_j) & \mathrm{SDRV}(C_jG_j) \\ \mathrm{SDRV}(C_jA_j) & \mathrm{SDRV}(A_jA_j) & \mathrm{SDRV}(A_jG_j) \\ \mathrm{SDRV}(C_jG_j) & \mathrm{SDRV}(A_jG_j) & \mathrm{SDRV}(G_jG_j) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{FDRV}(C_j) \\ \mathrm{FDRV}(A_j) \\ \mathrm{FDRV}(G_j) \end{bmatrix} \qquad [15]$$

where all first and second derivatives are evaluated at the stage *t* estimates of the item parameters.

An Earlier Approach to the Problem

In the bioassay setting, where the criterion (dosage level) is known, the preceding solution is correct. One approach to the psychometric setting, where the criterion (ability) is not known, is to replace the unknown ability parameters with provisional estimates. This approach is employed by computer programs such as LOGOG (Kolakowski & Bock, 1973), LOGIST (Wood, Wingersky, & Lord, 1977), and BICAL (Wright & Mead, 1978). LOGOG, for example, employs for the 2-parameter model an algorithm similar to one outlined below:

1. Use persons' logits of percent correct as provisional ability estimates.
2. Standardize provisional ability estimates.
3. On the basis of provisional ability estimates, form groups of persons with apparently similar abilities.
4. Assuming all persons in a group have the same true ability--the mean of their provisional estimates--solve Equations 4 and 5 to estimate item parameters.
5. Using provisional item parameter estimates, re-estimate person abilities.
6. Return to Step 2.

Cycles of this type were repeated until convergence was attained--which, it was learned, became less likely as the number of items and/or persons decreased. A major problem is the unreliability of the estimates of person ability when the number of items was small; in such cases, person ability estimates were a poor substitute for the true values.

Key Elements of the Bock-Aitkin Approach

An alternative does exist, however--an alternative that derives from long-standing procedures in the statistical literature in general and from an honorable tradition in psychometrics in particular (e.g., Kelley's paradox). The idea is this: Suppose that persons can be thought of as a random sample from a population in which ability is distributed in accordance with a distribution $g(\theta)$. Although each person's response vector x_i may not contain very much information about that person, it contains information about *g*. Taken together, the data of all persons may be sufficient to produce a fairly good characterization of *g*, which, in turn, may be used to condition and improve the inference about any individual person.

Now if *g* is a smooth distribution with finite moments, it may be approximated to any desired degree of accuracy by a discrete distribution over a finite number of points, i.e., a histogram. Let X_k, for k = 1, ..., q, be the points and let $A(X_k)$ be the densities at those points. By Bayes theorem, the posterior density of θ, given the response vector of person *i* is obtained as

$$P(X_k | \underset{\sim}{x}_i) = \frac{P(\underset{\sim}{x}_i | \theta_i = X_k)\ A(X_k)}{\sum_s P(\underset{\sim}{x}_i | \theta_i = X_s)\ A(X_s)} \qquad k=1, \ldots, q . \qquad [16]$$

Application to the estimation of item parameters is accomplished in the algorithm outlined below:

1. Using provisional estimates of item parameters, compute via Equation 1 the likelihood of each person's response pattern at each of the points, namely, $P(\underset{\sim}{x}_i | X_k)$.
2. Using given values (Bock & Aitkin, 1981) or provisional estimates (see below) of the densities $A(X_k)$ at each of the points, compute via Equation 16 the posterior probability that the ability of person _i_ is X_k.
3. (E-Step) Pseudo-counts of numbers of items attempted and number of items correct are then obtained by effectively distributing the data from each person over the points in proportion to the likelihood of his/her being there as follows:

$$N_{jk} = \sum_i d_{ij} \, P(X_k | \underset{\sim}{x}_i)$$

$$= \sum_i d_{ij} \frac{P(\underset{\sim}{x}_i | X_k) \, A(X_k)}{\sum_s P(\underset{\sim}{x}_i | X_s) \, A(X_s)} \qquad [17]$$

and

$$R_{jk} = \sum_i d_{ij} \, x_{ij} \, P(X_k | \underset{\sim}{x}_i)$$

$$= \sum_i d_{ij} \, x_{ij} \frac{P(\underset{\sim}{x}_i | X_k) \, A(X_k)}{\sum_s P(\underset{\sim}{x}_i | X_s) \, A(X_s)} \qquad [18]$$

where d_{ij} is 1 if person _i_ was presented item _j_ and 0 if not.

4. (M-step) The maximum likelihood equations for the item parameters, Equations 4 through 6, are then solved with respect to the pseudo-counts.
5. Unless item parameters are unchanged from the previous cycle, return to Step 1.

Bock and Aitkin (1981) showed that for given _g_, this procedure provides item parameter estimates that solve the marginal maximum likelihood equation

$$P(\text{data} | \text{item parameters}) = \prod_i P(\underset{\sim}{x}_i)$$

$$= \int_\Theta \prod_i P(\underset{\sim}{x}_i | \Theta) \, g(\Theta) \, d\Theta \qquad [19]$$

The problem with the "nuisance" ability parameters has been solved by integrating over their range, rather than by replacing them with estimates as in LOGOG or conditioning them away as is possible with the 1-parameter model only.

As a result, the unreliability in the ability estimate for a person has been ameliorated. Rather than basing the estimation of item parameters on a larger number of unreliable person ability estimates, they have been based on the much more stable estimates of population densities at various points along the ability scale and expected proportions of correct response at those points.

Extensions of the Bock-Aitkin Approach

The basic approach to estimating item parameters outlined above was shown by Bock and Aitkin to be a maximum likelihood solution under the conditions of (1) the 1- and 2-parameter normal ogive model, (2) all persons being administered the same set of items, and (3) the weights $A(X_k)$ remaining fixed throughout the solution, i.e., persons were in effect assumed to be a random sample from a known distribution. (By comparing item parameter estimates obtained with different priors on ability, this latter assumption was shown to be relatively unimportant; the item parameters varied little in the examples shown.) Since the publication of the article, progress has continued in the investigation of this approach. A number of extensions have been incorporated into the BILOG program.

Extension to the 3-parameter model. Along with the change to the logistic rather than to the normal ogive response curve, the provision for obtaining item parameter estimates in the 3-parameter model has been included. It is known that item parameter estimation in this model has been problematic. Certain improvement is achieved in the EM approach by the use of the estimation of provisional densities and probabilities at selected points rather than of person abilities, since proper estimates always exist for the former but not necessarily for the latter in the 3-parameter model. Difficulties remain, however, from another source: The matrix of second derivatives of the log likelihood function is often poorly conditioned in the 3-parameter model. The inversion of this matrix, required in the Newton-Raphson solution of Equations 4 through 6, can become unstable. This practical problem at least partly motivates the extension discussed immediately below.

Prior distributions on item parameters. In order to provide for stable and "reasonable" item parameter estimates in the 3-parameter model and in all models for small samples of persons, provision has been made for the incorporation of prior distributions on item parameters. For lower asymptotes, beta priors are employed; for slopes, log-normal; for intercepts, normal. (Priors are rarely necessary for intercepts; provision is made to facilitate linking studies, since the prior distribution of a given parameter may be based on a previous estimate and its standard error). The program provides Bayes modal estimates rather than maximum likelihood estimates when priors are used. Uncorrelated priors are assumed, thereby effecting a modification of the first derivatives Equations 4 through 6 by a so-called "penalty" function and the addition to the second double derivatives Equations 9, 12, and 14 of an augmenting term. The terms added to the diagonal of the matrix of second derivatives improve conditioning of this matrix. Solutions may be obtained from any data set with the imposition of sufficiently strong priors on the item parameters, though judicious and thoughtful choice of priors is recommended.

Estimation of the latent distribution. The original Bock-Aitkin solution assumes that persons are drawn from a specified distribution, normal or otherwise. The program now allows for the simultaneous estimation of the latent distribution if the user prefers. This is accomplished by revising the weights $A(X_k)$ at the beginning of each iteration as follows:

$$\begin{aligned} A^{(t+1)}(X_k) &= (1/N) \sum_i P^{(t)}(X_k \mid \underset{\sim}{x}_i) \\ &= \frac{1}{N} \sum_i \frac{P(\underset{\sim}{x}_i \mid X_k)\, A^{(t)}(X_k)}{\sum_s P(\underset{\sim}{x}_i \mid X_s)\, A^{(t)}(X_s)} \end{aligned} \qquad [20]$$

The distribution is then restandardized to set the scale and location of the latent ability variable. Under this convention, a common slope parameter is estimated in the 1-parameter model while the standard deviation of the latent distribution is fixed at one; this is equivalent to the more typical practice of fixing all slopes at one but not restricting the ability parameters.

Different patterns of item attempts for different persons. As seen in Equations 17 and 18, there is no necessity of assuming that all persons are presented the same items. This feature is of particular value in the assessment setting because item parameters may be estimated from data gathered in highly efficient multiple-matrix sampling designs where each person responds to only one to five items in a scale. Despite the sparsity of data for each person proscribing the estimation of his/her ability, it is no barrier to iteratively building up the estimates of population densities and item proportions correct at the points X_k. Persons with few responses are spread more broadly and persons with more responses are spread less broadly, each in accordance with the information conveyed by his/her response pattern.

Examples

In order to illustrate the use of the BILOG program, runs for 1-, 2-, and 3-parameter models are presented for two sets of data. First is a set of responses from 1,000 persons to five items of the Law School Admissions Test (LSAT), a data set which has been analyzed in the past by Bock and Lieberman (1970), Bock and Aitkin (1981), Andersen (1973), Andersen and Madsen (1977), and Thissen (1982). These data have been found to be well fit by a 1-parameter logistic item response model and a normal distribution of ability. Second is a set of simulated data of 1,000 persons to 18 items. The known parameters of the items, which include lower asymptotes, may be compared with the estimated values.

Example 1: LSAT

The five items of the LSAT analyzed by Bock and Lieberman in 1970 and others since were, on the whole, rather easy for the persons in the sample; about 30% of the examinees answered all five items correctly. It has been found by Andersen (1973) that the data are well fit by a 1-parameter logistic ogive model and an underlying normal distribution of ability. These data were subjected to item analysis via the 1-, 2-, and 3-parameter logistic models with BILOG, all under the assumption of an underlying normal distribution.

Table 1 presents the resulting item parameter estimates and, for the 1- and 2-parameter solutions, a likelihood ratio test of fit against a general multinomial alternative (see Bock & Aitkin, 1981). A straight maximum likelihood solution could not be obtained for the 3-parameter model, so the solution shown incorporates weak prior distributions on both slopes and asymptotes. The slopes had log normal prior distributions with means of zero (i.e., slopes of one) and standard deviations of two (slope values corresponding to a range of two standard deviations would be .018 and 54.598); asymptotes had a beta prior with parameters 1.25 and 5.75 (roughly comparable to saying with the weight of five observations that the asymptotes were .05). The formula for the likelihood ratio test was applied to the 3-parameter solution, but it must be noted that its distribution is not chi-square because the parameter estimates are modes of posteriors, not maximums of the likelihood function; its value, gauged in comparison with the degrees of freedom appropriate to a true maximum likelihood solution for the 3-parameter model, may be considered a somewhat more conservative index of fit.

Table 1
LSAT Item Parameter Estimates

Model	Chi-Square	df	Item	Threshold	Slope	Asymptote
1-P	9.90	19	1	-3.482	.788	.000
			2	-1.270	.788	.000
			3	-0.305	.788	.000
			4	-1.659	.788	.000
			5	-2.664	.788	.000
2-P	7.74	12	1	-3.318	.836	.000
			2	-1.356	.731	.000
			3	-0.279	.891	.000
			4	-1.845	.697	.000
			5	-3.074	.669	.000
3-P	9.27	7	1	-3.217	.831	.049
			2	-1.176	.752	.048
			3	-0.127	1.207	.029
			4	-1.704	.694	.048
			5	-3.114	.624	.050

It is no surprise to see that the 1-parameter model fits the data well and that the 2-parameter model fits even better but not sufficiently better to justify the additional parameters estimated. As noted by Thissen (1982), the 1-parameter solution agrees (after rescaling) with Andersen's conditional maximum likelihood solution (Andersen, 1973).

It is somewhat of a surprise to see that the 3-parameter solution appears to fit poorer than the 2-parameter solution, but this is because a maximum likelihood solution was not attained; the resulting parameter estimates depend not only on the data but on the priors. Bock and Lieberman (1970), estimating intercepts and slopes for different fixed values of asymptotes, found that asymptotes of zero did indeed fit best. It may be seen from the estimates of asymptotes that the only item which shows much difference from the prior is that of Item 3--the only item sufficiently difficult to provide much information about an asymptote. For this item, the information pushes the asymptote value down in the direction of zero.

Example 2: Simulated Data

Responses were generated from a random sample of 1,000 simulated examinees from a standard normal distribution to an 18-item test, in accordance with a 3-parameter logistic ogive item response model. The generating item parameters are shown in Table 2. There are essentially two groups of nine items each. In the first group, all slopes are 2.0 and all lower asymptotes are .05; thresholds range from -2.0 to +2.0 in increments of .5. In the second group, all slopes are 2.0 and all asymptotes are .25; thresholds again range from -2.0 to +2.0 in increments of .5. The broad range of difficulty of the items is reflected in their resulting proportion-correct values, which ranged from .11 to .96 correct. Item-test biserials ranged from .4 to .8.

Table 2
Generating Values of Item Parameters
for Simulated Data Example

Item	Threshold	Slope	Asymptote
1	-2.00	2.00	.05
2	-1.50	2.00	.05
3	-1.00	2.00	.05
4	-0.50	2.00	.05
5	0.00	2.00	.05
6	0.50	2.00	.05
7	1.00	2.00	.05
8	1.50	2.00	.05
9	2.00	2.00	.05
10	-2.00	2.00	.25
11	-1.50	2.00	.25
12	-1.00	2.00	.25
13	-0.50	2.00	.25
14	0.00	2.00	.25
15	0.50	2.00	.25
16	1.00	2.00	.25
17	1.50	2.00	.25
18	2.00	2.00	.25

BILOG solutions for the 1-, 2-, and 3-parameter models are shown in Table 3. The 1- and 2-parameter solutions are straight maximum likelihood solutions, with the normal distribution of persons assumed. The 3-parameter solution required priors on all item parameters, the specification of which is described in

Table 3
Item Parameter Estimates for Simulated Data
for the 1-, 2-, and 3-Parameter Models

Item	Intercept	SE	Slope	SE	Threshold	SE	Dispersion	SE	Asymptote	SE	Chi-Square	df	Prob
1-Parameter Model													
1	-3.632	.141	1.197	.015	-3.035	.141	.836	.011	.0	.0	7.7	9	.5640
2	-3.324	.117	1.197	.015	-2.777	.117	.836	.011	.0	.0	26.3	9	.0019
3	-2.083	.089	1.197	.015	-1.741	.089	.836	.011	.0	.0	29.3	9	.0006
4	-1.415	.081	1.197	.015	-1.182	.082	.836	.011	.0	.0	46.1	9	.0000
5	-0.384	.074	1.197	.015	0.320	.075	.836	.011	.0	.0	20.3	9	.0161
6	0.391	.078	1.197	.015	0.327	.079	.836	.011	.0	.0	39.8	9	.0000
7	1.272	.084	1.197	.015	1.063	.085	.836	.011	.0	.0	25.7	9	.0024
8	1.885	.095	1.197	.015	1.575	.096	.836	.011	.0	.0	8.5	9	.4840
9	2.196	.105	1.197	.015	1.835	.105	.836	.011	.0	.0	29.9	9	.0005
10	-4.603	.170	1.197	.015	-3.847	.170	.836	.011	.0	.0	30.0	9	.0005
11	-2.867	.109	1.197	.015	-2.396	.110	.836	.011	.0	.0	4.2	9	.8961
12	-2.619	.100	1.197	.015	-2.188	.101	.836	.011	.0	.0	23.9	9	.0046
13	-1.616	.081	1.197	.015	-1.350	.082	.836	.011	.0	.0	19.7	9	.0196
14	-0.818	.072	1.197	.015	-0.684	.073	.836	.011	.0	.0	20.7	9	.0142
15	-0.301	.070	1.197	.015	-0.251	.070	.836	.011	.0	.0	26.5	9	.0018
16	0.275	.071	1.197	.015	0.230	.072	.836	.011	.0	.0	28.6	9	.0008
17	0.669	.071	1.197	.015	0.559	.072	.836	.011	.0	.0	57.1	9	.0000
18	0.837	.073	1.197	.015	0.700	.073	.836	.011	.0	.0	56.8	9	.0000
All Items											501.5	162	.0000
2-Parameter Model													
1	-3.587	.142	1.515	.116	-2.368	.149	.660	.050	.0	.0	3.7	8	.8821
2	-3.341	.121	2.008	.105	-1.664	.121	.498	.026	.0	.0	8.5	8	.3370
3	-1.982	.092	1.922	.105	-1.032	.095	.520	.029	.0	.0	13.5	8	.0958
4	-1.332	.087	2.168	.122	-0.614	.089	.461	.026	.0	.0	19.5	8	.0123
5	-0.154	.075	1.490	.103	-0.104	.088	.672	.046	.0	.0	5.6	8	.6933
6	0.695	.081	1.747	.113	0.398	.088	.573	.037	.0	.0	19.2	8	.0141
7	1.522	.085	1.368	.092	1.113	.097	.732	.049	.0	.0	22.7	8	.0038
8	2.111	.096	1.190	.089	1.774	.113	.840	.063	.0	.0	14.6	8	.0673
9	2.275	.103	0.735	.092	3.093	.199	.360	.170	.0	.0	13.4	8	.0995
10	-4.806	.175	2.235	.126	-2.149	.173	.447	.026	.0	.0	12.3	8	.1356
11	-2.657	.109	1.318	.101	-2.016	.122	.758	.058	.0	.0	4.5	8	.8095
12	-2.487	.102	1.562	.098	-1.593	.108	.640	.040	.0	.0	15.2	8	.0543
13	-1.418	.081	1.333	.090	-1.150	.099	.811	.060	.0	.0	12.0	8	.1510
14	-0.625	.072	1.034	.085	-0.604	.106	.967	.080	.0	.0	15.1	8	.0566
15	-0.117	.068	.885	.079	-0.133	.122	1.130	.101	.0	.0	10.9	8	.2067
16	0.446	.070	.944	.081	0.473	.114	1.059	.091	.0	.0	20.8	8	.0079
17	0.761	.069	.501	.071	1.521	.291	1.997	.283	.0	.0	15.2	8	.0558
18	0.918	.071	.466	.072	1.971	.338	2.148	.330	.0	.0	12.6	8	.1250
All Items											239.3	144	.0000
3-Parameter Model													
1	-3.363	.367	1.328	.167	-2.532	.280	.753	.094	.053	.032	7.6	7	.3729
2	-3.232	.407	1.956	.192	-1.652	.275	.512	.050	.050	.030	6.8	7	.4545
3	-1.804	.332	1.795	.138	-1.005	.276	.557	.043	.052	.030	6.7	7	.4642
4	-1.131	.346	1.936	.138	-0.584	.303	.517	.037	.035	.021	19.3	7	.0074
5	-0.027	.348	1.463	.123	-0.018	.338	.683	.057	.040	.023	5.1	7	.6537
6	0.852	.454	1.803	.154	0.472	.450	.554	.047	.035	.016	9.3	7	.2278
7	1.991	.551	1.832	.167	1.087	.577	.546	.050	.038	.013	11.1	7	.1313
8	2.588	.643	1.542	.183	1.679	.692	.649	.077	.030	.011	5.8	7	.5592
9	3.089	.814	1.266	.215	2.439	.901	.790	.134	.039	.013	8.1	7	.3277
10	-4.325	.481	1.858	.233	-2.327	.301	.538	.067	.052	.032	14.7	7	.0400
11	-2.571	.308	1.283	.135	-2.004	.249	.779	.082	.051	.031	2.2	7	.9443
12	-2.217	.376	1.605	.157	-1.382	.316	.623	.061	.186	.051	7.0	7	.4317
13	-1.101	.404	1.442	.145	-0.764	.390	.693	.070	.196	.052	18.3	7	.0110
14	-0.125	.561	1.465	.183	-0.086	.573	.682	.086	.214	.048	14.1	7	.0484
15	0.484	.545	1.432	.171	0.338	.565	.698	.083	.192	.037	6.7	7	.4570
16	1.274	.613	1.765	.188	0.722	.637	.567	.061	.156	.026	8.3	7	.3080
17	1.670	.680	1.009	.180	1.654	.764	.990	.176	.164	.033	18.6	7	.0096
18	2.093	.792	1.193	.210	1.756	.877	.839	.148	.165	.027	11.2	7	.1277
All Items											181.0	126	.0010

greater detail below. The indices of goodness of fit that accompany the estimates are not true likelihood chi-squares, but approximations based on combining persons into 10 homogeneous groups on the basis of their Bayes ability estimates. Counts of correct responses observed in each group were then compared with those expected under the assumption that all persons in a group have the same true ability.

The 1-parameter solution exhibits biases in both thresholds and slopes, as compared with the generating values. Although all items have the same generating slope of 2.0, the common value estimated is only 1.2, due to the attenuation caused by the nonzero lower asymptotes. There is a tendency for difficult items to fit more poorly than easy items, and for items of the second group (with higher asymptotes) to fit more poorly than items of the first group (with lower asymptotes).

The 2-parameter solution represents a marked improvement in fit. Many items, particularly easier items, are well explained by this solution. Serious biases are apparent, however, in the slope estimates. Again, because of the nonzero lower asymptotes, slopes are consistently underestimated to a degree that increases with difficulty and with the values of the asymptote itself.

The 3-parameter solution represents another, though less impressive, improvement in fit; the solution required prior distributions on intercepts, slopes, and asymptotes. Normal priors with mean zero and standard deviation two were placed on all intercepts. It may be seen that the data effectively dominated the prior in this case, as considerable information about intercepts is present in the data. Log-normal priors with mean .588 and standard deviation .500 were placed on slopes; this corresponds roughly to a prior mean of 1.8 and a standard deviation of 1.0 for the slopes themselves, suggesting a prior belief that slopes would probably range between about .5 and 5.0. Beta priors with parameters (3.5, 47.5) were placed on asymptotes for the first 11 items and with parameters (11, 41) for the last 7 items; this corresponds to saying with the weight of 50 observations that the asymptotes were .05 for the first 11 items and .20 for the last 7. These values were obtained by inspecting plots of the residuals from the 2-parameter solution, as illustrated by Figure 1.

Although the 3-parameter solution provides an adequate fit to the data, with a chi-square ratio less than one and a half, discrepancies remain between final parameter estimates and generating values. For the second group of items in particular, both thresholds and slopes tend to be too low. The apparent paradox of adequate fit but imperfect recovery of item parameters is resolved at least partially by an examination of estimated and observed response curves. Figure 1 plots data for Item 16 under the 2-parameter solution; Figure 2 plots the data for the same item under the 3-parameter solution. Despite nontrivial differences in estimates of item parameters (.5 vs. .7 for threshold, .9 vs. 1.8 for slope, 0 vs. .16 for asymptote), both curves are able to explain observed proportions of correct response in the region where the majority of persons are to be found. Despite the differences in their parameters, the 2- and 3-parameter curves are not very different with respect to the data at hand.

Figure 1
Observed and Expected 2-Parameter Logistic Response Curve for Item 16
(Smooth Line is Fitted Response Curve; "X" Represents Proportion Correct of a Group of Persons with Approximately Similar Abilities; Vertical Bars around Curve Represent Two Standard Errors around Expected Group Proportions Correct)

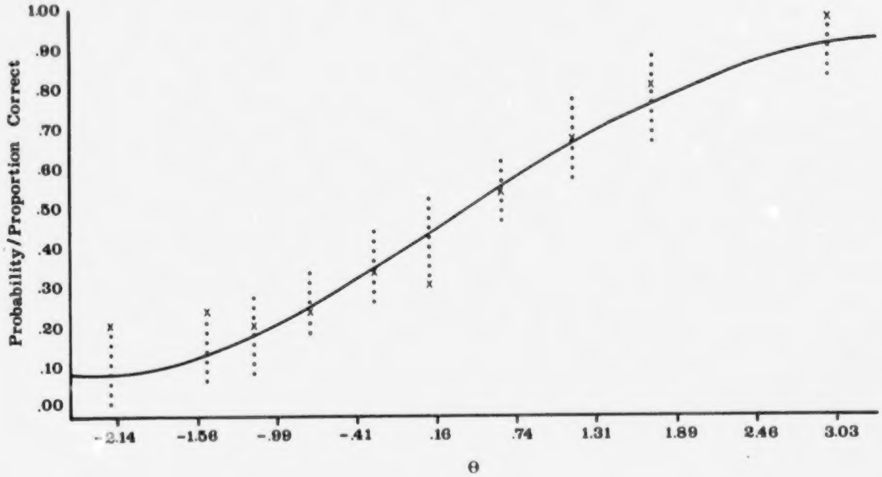
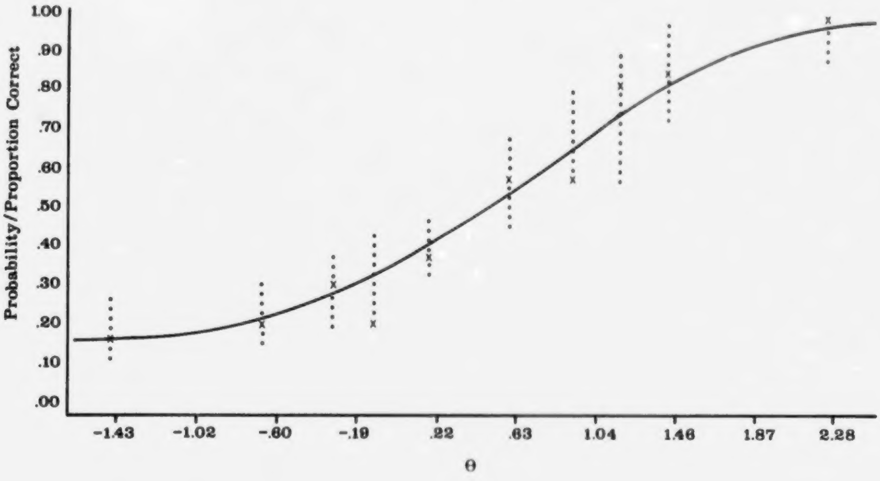


Figure 2
Observed and Expected 3-Parameter Logistic Response Curve for Item 16
(Smooth Line is Fitted Response Curve; "X" Represents Proportion Correct of a Group of Persons with Approximately Similar Abilities; Vertical Bars around Curve Represent Two Standard Errors around Expected Group Proportions Correct)



Discussion

With the use of marginal maximum likelihood estimation procedures and prior distributions on item parameters, it is now possible to estimate item response curves under the 1-, 2-, and 3-parameter logistic models from even very sparse data sets. It will be noted that the emphasis here is on the estimation of response curves rather than on item parameters. Simulation studies suggest that the recovery of generating item parameters is problematic, even with large numbers of items and persons, when the parameters of an item are not well identified by the calibration sample. These circumstances seem to obtain quite frequently with the 3-parameter model and, occasionally, with the 2-parameter model when the calibration sample does not span a sufficiently broad range of ability. Item response curves are estimated that do, on the other hand, explain the data satisfactorily.

The explanation of these findings is that for typical educational tests, data are well explained by a region of values in the parameter space. For an easy item, for example, data at hand may be well explained by either a 2- or a 3-parameter ogive; curves of each type can be found that are virtually identical in the region of the ability scale where the calibration examinees are to be found. The use of weak prior distributions will function in this situation to keep the resulting parameter estimates "reasonable," or in line with the values that the substantive interpretations of the item parameters would suggest (e.g., item slopes ranging between, say, 0 and 4) and asymptotes ranging between, say, 0 and .25).

The practical implication of this result is that the substantive interpretation of item parameters in the 3-parameter model (and, to a lesser extent, the 2-parameter model as well) may not always be justified. Maximum likelihood estimates for a given item may differ substantially from another set of values that reproduce the calibration data nearly as well. Discussion of item characteristics could be couched in terms of the item information function instead, since all sets of item parameter estimates in the "solution space" will yield similar information functions in the region where the data lies. Characteristics such as the point of maximum information and the value of the information function at that point can be expected to be much more stable than the item parameter estimates themselves.

Fortunately, most applications of IRT depend on the shape and location of response curves rather than the parameter values, particularly when applications are foreseen for examinees who are typical of the calibration sample. The estimation of an individual's ability from a given response pattern would typically be similar if computed from any item parameter values that produce similar response curves in the neighborhood of his/her ability. Discrepancies would be more likely for persons with abilities that are extreme.

One application that demands special attention, however, is vertical equating, or the linking of tests across broad ranges of ability--often across several grades or age groups. One approach to the equating problem is to calibrate tests separately in the low and high ability groups, say, and then to attempt to find the linear transformation that produces the closest match of item parameter estimates for those items that were administered to both groups. Now, a linking item will tend to be comparatively easy for the high ability group and comparatively difficult for the low ability group. This means that the range of ability for which its response curve is well estimated in either group does not cover the region where the groups overlap, i.e., where the two estimated curves are supposed to be made to match. Poor linking may result as an artifact of the multicollinearity of item parameter estimates. The information needed for a proper link is found in not just the item parameter estimates and their standard errors, but in the matrix of correlations among the estimates as well. (This problem may be avoided by calibrating all items together with responses from all groups simultaneously, an option available in both BILOG and LOGIST.)

REFERENCES

Andersen, E. B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, *38*, 123-140.

Andersen, E. B., & Madsen, M. Estimating the parameters of a latent population distribution. *Psychometrika*, 1977, *42*, 357-374.

Bock, R. D., & Aitkin, M. Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 1981, *46*, 443-459.

Bock, R. D., & Lieberman, M. Fitting a response model for *n* dichotomously scored items. *Psychometrika*, 1970, *35*, 179-197.

Bock, R. D., & Mislevy, R. J. *BILOG: Maximum likelihood item analysis and test scoring; binary logistic models*. Chicago: International Educational Services, 1982.

Dempster, A. P., Laird, N., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society*, *Series B*, 1977, *39*, 1-38.

Kolakowski, D., & Bock R. D. *LOGOG: Maximum likelihood test scoring and item analysis; logistic model for multiple item responses*. Chicago: International Educational Services, 1973.

Thissen, D. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 1982, *47*, 175-186.

Wood, R. L., Wingersky, M., & Lord, F. M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (ETS RM-76-6). Princeton NJ: Educational Testing Service, 1976.

Wright, B. D., & Mean, R. *BICAL: Calibrating items and scales with the Rasch model*. Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1978.

DISCUSSION

CHARLES LEWIS
UNIVERSITY OF GRONINGEN

I have enjoyed very much hearing, reading, and thinking about the two papers on estimation with the EM algorithm by Tsutakawa and by Mislevy and Bock. This is not primarily because of the common EM theme--although the approach seems very promising in its application to IRT--but because a "random effects" model for latent ability is adopted in both papers. The idea that ability should be treated as a random variable is one of the foundation stones of classical test theory but has been neglected in the development of modern IRT. (The writings of Bock and his colleagues form the most important exception to this generalization.)

A probable cause for this state of affairs may be traced to the goal of some of the developers of IRT to obtain "person-free" item measurement, that is, to be able to describe characteristics of items that are independent of the abilities of the persons answering the items. Adopting a particular family of distributions (such as the normal distribution) to describe populations of abilities may have been considered to be counterproductive to this goal. In fact, there is good evidence that recognition of abilities as random effects may, in many cases, substantially improve statistical inferences regarding both item characteristics and the abilities themselves, as well as characteristics of the population distribution of abilities. Moreover, the actual definitions of item parameters (and abilities) are made independently of an assumed ability population, so these parameters remain "person-free" (and, concomitantly, the abilities remain "item-free").

Tsutakawa's paper--developing themes introduced by Rigdon and Tsutakawa (1981) and, of course, the basic EM theme of Dempster, Laird and Rubin (1977)--considers primarily theoretical issues. Thus, he shows that if the general EM (GEM) algorithm converges, the resulting parameter estimates satisfy the original likelihood equations. This is, of course, neither a guarantee of convergence of the algorithm nor a proof that the GEM estimates (given convergence) are maximum likelihood estimates, and Tsutakawa explicitly acknowledges this. Nonetheless, it is an important minimal condition for the procedure that is discussed.

In the final section of his paper, Tsutakawa shows that for the Rasch model in its logistic form, with normally distributed ability, if the GEM converges to a finite solution, then the estimates so obtained will maximize the likelihood. What is not guaranteed in this case, contrary to the last sentence in the introductory section, is the actual convergence of the GEM algorithm. Again, it is an important result as far as it goes, but the complete story has not yet been told, even for this "simple" case.

For the "fixed" abilities case of the Rasch model, Fischer (1981) has provided necessary and sufficient conditions for the existence and uniqueness of solutions to the likelihood equations for both the unconditional case (where abilities are explicitly estimated) and the conditional case (where abilities do not appear, due to conditioning on number-correct scores). Even in this paper, however, actual convergence of a particular algorithm for solving the likelihood equations is not proven.

To illustrate the GEM procedure, Tsutakawa turns to the 2-parameter logistic model, with normally distributed abilities, and obtains equations parallel to, but much simpler than, the corresponding likelihood equations. For k items, the GEM algorithm leads to repeatedly solving *k pairs* of equations, each with only two unknowns, whereas with the likelihood equations, 2k equations in 2k unknowns must be solved at one time.

One remaining difficulty with *either* approach is that, at each iteration, numerical integrations with respect to the posterior distribution of ability, given a person's responses, must apparently be carried out for each of the n persons whose responses are being analyzed. (See Tsutakawa's Equations 25, 26, 28, and 29, where posterior expectations appear within summations with respect to the person index i.) I shall return to this point when comparing the algorithms proposed in the two papers.

Turning now to the contribution of Mislevy and Bock, the themes they develop appear to arise primarily from Bock and Aitkin (1981). As suggested by its title, "Implementation of the E-M Algorithm in BILOG," Mislevy and Bock's work is more "practically" oriented than Tsutakawa's, which is to say that they have given no attention to conditions under which a maximum likelihood solution might be obtained by their algorithm. (In all fairness, it should be said that Bock and Aitkin had already considered this issue for a very similar model: the 2-parameter normal ogive.) Instead, Mislevy and Bock are more explicit about how to solve their equations and how to carry out numerical integrations. They also deal with a more complex model than Tsutakawa's (the 3-parameter logistic), they explicitly include the possibility of analyzing incomplete response designs, and they allow more general classes of population ability distributions than the normal distribution, including general discrete distributions whose probability values are estimated simultaneously with the item parameters.

In addition to these extensions, Mislevy and Bock discuss the possibility of using prior densities for item parameters and, after suitably modifying their algorithm, obtaining posterior modal *Bayesian* estimates of these parameters. Although they seem to view this option primarily as a means of alleviating possible numerical problems in the estimation process, a potential user should realize that a fundamentally different philosophy of statistical inference has been adopted here than that on which maximum likelihood is based. It happens to be a philosophy which I find very appealing, and I hope this aspect of their work is further developed in the future.

It is perhaps worthwhile noting here that there is nothing "Bayesian" (in the sense of combining prior knowledge with data) about treating ability as a random variable. To the extent that the group of persons whose responses are analyzed may be regarded as a random sample from a fixed population, analyzing their abilities as random effects has a classical sampling theory justification. A complete Bayesian treatment of abilities would require the adoption of priors for the characteristics of the ability distribution. This approach has been developed for linear models by Lindley and Smith (1972), and for the Rasch item response model by Leonard (1972), by Jansen (1981), and by Swaminathan and Gifford (1982).

As mentioned earlier in the discussion of Mislevy and Bock's paper, when working with a general discrete distribution for ability, estimates of the population probabilities associated with the possible discrete values are obtained, using Equation 20 iteratively. Although they do not say as much, it may be shown that the estimates obtained, together with the estimated item parameters (assuming that the combined algorithm converges), satisfy the likelihood equations for both sets of parameters and thus, *possibly* are joint maximum likelihood estimators for the ability distribution and item parameters.

A minor point regarding Mislevy and Bock's explanation of the use of Equation 20 iteratively concerns restandardization. In the case of a discrete distribution, standardizing implies changing, not the probabilities, but the set of possible ability values, to guarantee a mean of zero and standard deviation of unity. It should be emphasized that this is *not* necessary for identifying the item parameters, as would be the case, for example, in working with a normal distribution. Once the possible ability values have been specified, the origin and the unit of the ability scale are fixed as well, and there is no identification problem for the item parameters. Restandardization *may*, of course, be carried out, either at each iteration or after convergence has been obtained, but it should not be forgotten that the corresponding transformation *must* also be carried out on the item slopes and intercepts to guarantee that (in their notation) the values of $A_j X_k + C_j$ do not change. Otherwise, restandarization would destroy the (relative) optimality of the solution obtained before standardizing. Though Mislevy and Bock do not say as much, it must be assumed that they actually carry out this compensating item parameter transformation at each iteration.

Their paper concludes with the presentation of two sets of analyses--one of real data (five LSAT items) and the other of simulated data. Their discussion of these examples is thoughtful, though the results are, in two respects, less than what might have been hoped. First, the inability of either example to obtain maximum likelihood estimates for the 3-parameter model must be considered a disappointment. That Bayesian modal estimates can be found is scant comfort to those preferring the sampling theory approach to statistical inference.

Second, the instability of parameter estimates across models in the first example (where all three models fit well), and the inaccuracy of the estimates for the 3-parameter model (on which the simulated data were based) in the second example, is a distressing phenomemon. Why, for instance, should the estimated slope for Item 3 in the first example (Table 1) be so much steeper in the 3-parameter solution than in the 2-parameter solution? In the second example, why should *all* the estimated slopes for the 3-parameter solution (Table 3) be below the theoretical values (Table 2)? Mislevy and Bock provide a partial answer by stressing (1) that the item response curves obtained from widely varying parameter values may be almost invariant and (2) that it is these curves that are of primary importance in most (though not all) applications of IRT.

One minor frustration that I had with this section concerns the goodness-of-fit tests presented for the first example. Reference is made to Bock and Aitkin (1981) for an explanation of the test. Nothing, however, is said about the discrepancy (both in chi-square value and in degrees of freedom) between the test reported for the 1-parameter model in Table 1 and what appears to be the same test of the same model for the same data given in Table 3 of Bock and Aitkin (1981, p. 455). No doubt there is a simple explanation (such as grouping of response patterns with small expected frequencies in one case but not in the other). It would just be nice to know what it is!

Finally, I would like to make a comparison of the algorithms presented by Tsutakawa and by Mislevy and Bock. They appear to be two different generalizations of the basic EM algorithm; and, indeed, Bock and Aitkin (1981, p. 448) claim that what they propose "is not the same as the general EM algorithm" (i.e., that discussed by Tsutakawa). In fact, the GEM can be used to generate the EM equations given by Bock and Aitkin (1981) and by Mislevy and Bock as special cases. I will sketch a proof of this result for the 3-parameter logistic model with a general discrete distribution of ability, no priors on the item parameters, and with a complete set of responses from all persons to all items. (This is the simplest situation discussed by Mislevy and Bock.)

Based on the general item response model described by Tsutakawa in his third section (not the 2-parameter model discussed in the fourth section), the crucial function Q from his Equation 2 can be written (following his notation as closely as possible) as

$$Q(\underset{\sim}{\beta}',\underset{\sim}{\lambda}'|\underset{\sim}{\beta},\underset{\sim}{\lambda}) = E\{\sum_{ij}\sum[y_{ij}\log p'_{ij} + (1 - y_{ij})\log(1 - p'_{ij})] + \sum_{i} \log p(\theta_i|\underset{\sim}{\lambda}')|\underset{\sim}{y},\underset{\sim}{\beta},\underset{\sim}{\lambda}\} \quad . \qquad [1]$$

Here, p'_{ij} is used to denote the probability of a positive response to an item with parameters $\underset{\sim}{\beta}'_J$ for a person with ability θ_i. (This expression is analogous to Tsutakawa's Equation 27 for the 2-parameter model.)

Taking derivatives with respect to $\underset{\sim}{\beta}'_J$ and taking the expectation of the resulting sum term by term gives

$$\frac{\partial Q}{\partial \underset{\sim}{\beta}'_J} = \sum_{i} E\left[\left(\frac{y_{iJ} - p'_{iJ}}{p'_{iJ}(1-p'_{iJ})}\right)\left(\frac{\partial p'_{iJ}}{\partial \underset{\sim}{\beta}'_J}\right) \middle| \, \underset{\sim}{y}_i \, , \underset{\sim}{\beta}, \underset{\sim}{\lambda}\right] \quad . \qquad [2]$$

This is essentially the form used by Tsutakawa to obtain Equations 28 and 29.

Now suppose that the discrete values which ability may assume are

$$\theta_{(k)} \ , \ k = 1\ (1)\ q \ . \qquad [3]$$

If the posterior probability that θ_i takes on the value $\theta_{(k)}$, given the response pattern $\underset{\sim}{y}_i$, is denoted by

$$p(\theta_{(k)} \mid \underset{\sim}{y}_i \ , \ \underset{\sim}{\beta} \ , \ \underset{\sim}{\lambda}\ ) \ , \qquad [4]$$

the value of p'_{iJ} when θ_i equals $\theta_{(k)}$ by $p'_{(k)J}$ and the notation

$$\underset{\sim}{W}'_{(k)J} = \frac{1}{p'_{(k)J}(1-p'_{(k)J})} \ \frac{\partial p'_{(k)J}}{\partial \underset{\sim}{\beta}'_J} \ , \qquad [5]$$

is used, then the derivative in Equation 2 may be rewritten as

$$\frac{\partial Q}{\partial \underset{\sim}{\beta}'_J} = \sum_{ik}\sum [(y_{iJ} - p'_{(k)J})\ \underset{\sim}{W}'_{(k)J}\ p(\theta_{(k)} \mid \underset{\sim}{y}_i \ , \ \underset{\sim}{\beta},\ \underset{\sim}{\lambda})] \ . \qquad [6]$$

Reversing the order of summation and letting (with notation analogous to that used by Mislevy and Bock in their Equations 17 and 18)

$$R_{(k)J} = \sum_i y_{iJ}\ p(\theta_{(k)} \mid \underset{\sim}{y}_i \ , \ \underset{\sim}{\beta},\ \underset{\sim}{\lambda}) \ , \qquad [7]$$

and

$$N_{(k)J} = \sum_i p(\theta_{(k)} \mid \underset{\sim}{y}_i \ , \ \underset{\sim}{\beta},\ \underset{\sim}{\lambda}) \ , \qquad [8]$$

gives

$$\frac{\partial Q}{\partial \underset{\sim}{\beta}'_J} = \sum_k (\ R_{(k)J} - p'_{(k)J} N_{(k)J}\ )\ \underset{\sim}{W}'_{(k)J} \ . \qquad [9]$$

From this expression, Mislevy and Bock's Equations 4, 5, and 6 are easily derived, assuming $p'_{(k)J}$ is given by

$$p'_{(k)J} = G'_J + (1 - G'_J)\ \Psi(A'_J\ \theta_{(k)} + C'_J) \ , \qquad [10]$$

and

$$\underset{\sim}{\beta}'_J = \begin{pmatrix} C'_J \\ A'_J \\ G'_J \end{pmatrix} . \qquad [11]$$

In a similar vein, Mislevy and Bock's Equation 20 may be obtained for the parameters of the ability distribution. Begin by defining

$$\lambda_{(k)} = \mathrm{Prob}(\theta_i = \theta_{(k)} \mid \underset{\sim}{\lambda}\ ) \ , \ k = 1(1)q \ . \qquad [12]$$

Returning to the expression in (my) Equation 1 for the basic function Q, the second term may be rewritten as

$$\sum_{i}\sum_{k} \log \lambda'_{(k)}\, p(\theta_{(k)} \mid \underset{\sim}{y}_i , \underset{\sim}{\beta}, \underset{\sim}{\lambda}) \quad . \qquad [13]$$

Using the Lagrange multiplier μ to incorporate the side condition

$$\sum_{k} \lambda'_{(k)} = 1 \qquad [14]$$

and setting the derivatives of the resulting expression with respect to $\lambda'_{(K)}$, equal to zero gives the equations

$$\frac{1}{\lambda'_{(K)}} \sum_{i} p(\theta_{(K)} \mid \underset{\sim}{y}_i , \underset{\sim}{\beta}, \underset{\sim}{\lambda}) - \mu = 0 \quad . \qquad [15]$$

Solving Equation 15 for $\lambda'_{(K)}$ gives

$$\lambda'_{(K)} = \frac{\sum_{i} p(\theta_{(K)} \mid \underset{\sim}{y}_i , \underset{\sim}{\beta}, \underset{\sim}{\lambda}) \quad ,}{\mu} \qquad [16]$$

and summing over K reveals that

$$\mu = n \quad , \qquad [17]$$

the total number of persons. Except for changes in notation, my Equation 16 is the same as Equation 20 of Mislevy and Bock.

Thus, at least one special application of Bock and Aitkin's general approach can be shown to be a special case of the GEM as well. For other cases, the proof would follow similar lines. Continuous ability distributions should be approximated by discrete ones, as Mislevy and Bock state just above their Equation 16. This enables avoiding the problem alluded to earlier in my discussion of Tsutakawa's results--that of having to carry out numerical integrations for each person. Instead, summations are carried out over persons first for each discrete ability value (see my Equations 7 and 8), followed by summation over the possible abilities (see my Equation 9). Without such a simplification, the practical applications of the GEM algorithm would be limited to relatively small samples of persons.

These two papers complement each other very well and together bring our understanding of how to work with abilities as random effects in IRT a considerable step further.

References

Bock, R. D., & Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 1981, 46, 443-459.

Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B*, 1977, *39*, 1-38.

Fischer, G. H. On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, 1981, *46*, 59-77.

Jansen, G. G. H. *A Bayesian latent trait model for binary test items*. Paper presented at the annual meeting of the European Mathematical Psychology Society, Birmingham, England, 1981.

Leonard, T. *Bayesian methods for discrete data*. Unpublished manuscript, Iowa City, IA, 1972.

Lindley, D. V., & Smith, A. F. M. Bayesian estimates for the linear model (with Discussion). *Journal of the Royal Statistical Society, Series B*, 1972, *34*, 1-41.

Rigdon, S. E., & Tsutakawa, R. K. *Estimation in latent trait models* (Research Report 81-1; Mathematical Sciences Technical Report No. 102). Columbia, MO: University of Missouri, Department of Statistics, 1981.

Swaminathan, H., & Gifford, J. A. Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 1982, *7*, 175-191.

A STATISTICAL PROCEDURE FOR ASSESSING TEST DIMENSIONALITY

WILLIAM STOUT
UNIVERSITY OF ILLINOIS

An important statistical problem in psychological test theory is the development of a sound method for determining whether a test which purports to measure the level of a certain ability is, in reality, significantly contaminated by the varying levels of one or more other abilities displayed by persons taking the test. For example, is a test of mathematical ability contaminated by varying levels of verbal ability displayed by persons taking the test or is a test of reading ability contaminated by varying levels of familiarity with middle-class American culture displayed by persons taking the test? Because of the large number of private and governmental organizations routinely using tests to screen people for the levels of various abilities, this problem of assessing the dimensionality of a test is of great importance.

The solution will be useful in settings other than psychological testing, since the problem is one of general interest and should, hence, be an important addition to the statistical methodology literature. Thus, it seems appropriate now to give a careful abstract statement of the problem, independent of its psychometric context.

Consider sampling units from a population and applying several treatments to each sampled unit. Suppose that the outcome of each unit-treatment combination is either success or failure. Suppose that associated with each unit is a parameter, $\underset{\sim}{\theta}$ (the ability parameter), which determines the likelihood of each treatment being successful for that unit. Assume that the dimensionality of $\underset{\sim}{\theta}$ is unknown (the precise mathematical definition of the dimensionality of $\underset{\sim}{\theta}$ will be given below). Thus, for each unit, dichotomous random variables $\{U_i\}$ are observed, where *i* is the treatment index. Let "treatment characteristic curves" $\{P_i(\cdot)\}$ be defined by

$$P_i(\underset{\sim}{\theta}) = P[U_i = 1|\underset{\sim}{\Theta} = \underset{\sim}{\theta}] = 1 - P[U_i = 0|\underset{\sim}{\Theta} = \underset{\sim}{\theta}] \;, \qquad [1]$$

the probability of treatment *i* being successful, given that the sampled unit has ability θ. It is assumed that the process of random sampling units induces a probability distribution on the population of units with associated random variable $\underset{\sim}{\Theta}$.

Purpose

Described in this paper is an approach to the problem of finding a theoretically sound and useful procedure for making inferences about the dimensionality of $\underline{\theta}$, that is, more precisely, the dimensionality of the distribution of $\underline{\Theta}$. In order that this problem be well formulated mathematically, the dimensionality of $\underline{\Theta}$ needs to be defined precisely. The definition (Levine, 1981) that is used depends on the asymptotic behavior of "formula sequences." To define a linear formula sequence, a linear formula score must first be defined.

Definition of a Linear Formula Score

Given the outcomes $(U_1, U_2, \ldots, U_n)$ of *n* treatments resulting from a sampled unit, a linear formula score is a score of the form

$$\alpha_n = \sum_{i=1}^{n} a_i^{(n)} U_i \qquad [2]$$

provided that

$$a_i^{(n)} \geq 0, \ \sum_{i=1}^{n} a_i^{(n)} = 1 \ . \qquad [3]$$

Then, a formula sequence is a sequence of linear formula scores $(\alpha_1, \alpha_2, \ldots, \alpha_n, \ldots)$ such that, referring to Equation 2,

$$a_i^{(n)} \ a_{i'}^{(n+1)} = a_i^{(n+1)} \ a_{i'}^{(n)} \qquad [4]$$

for all $i' \leq n$, $i \leq n$, and $n \geq 1$. The content of Equation 4 is that the contribution of a treatment, say, *i*, relative to another treatment, say, *i'*, is the same for all linear formula scores α_n for which $n \geq i$, $n \geq i'$. The prototype of a linear formula score and a formula sequence is the proportion-correct

$$\sum_{i=1}^{n} U_i/n \qquad [5]$$

and

$$\{U_1, (U_1 + U_2)/2, \ldots, \sum_{i=1}^{n} U_i/n, \ldots\} \qquad [6]$$

respectively. Levine's (1981) definition can now be stated (below, $\mathrm{Var}\,[X|Y]$ denotes the variance of X, given Y):

> A sequence of dichotomous random variables $\{U_1, U_2, \ldots, U_n, \ldots\}$ is *d* dimensional if there exist *d* formula sequences $\{h_1^{(n)}\}$, $\{h_2^{(n)}\}$, ..., $\{h_d^{(n)}\}$ such that for every formula sequence $\{h^{(n)}\}$,

$$\mathrm{Var}\ [h^{(n)} | h_1^{(n)}, \ldots, h_d^{(n)}] \to 0 \qquad [7]$$

as $n \to \infty$; and, moreover, no smaller d works.

Note that it is the set of observables $\{U_1, U_2, \ldots, U_n, \ldots\}$ that is d dimensional. The ability θ is not observable and is known only by inference. Nonetheless, let it be said that $\underset{\sim}{\theta}$ is d dimensional, meaning that a d-dimensional random vector $\underset{\sim}{\theta}$ and treatment characteristic curves $\{P_i(\cdot)\}$ (the conditional distributions of the U_i's given $\underset{\sim}{\theta}$) can be constructed to specify the joint probability law of the d-dimensional U_i's.

Assessment of Test Dimensionality

As stated above, the dimensionality problem is of particular importance in the field of psychological testing. In this case, the units are persons and the treatments are test items. The function $P_i(\cdot)$ is called the item characteristic curve for the ith item. The administration of a psychological test is modeled as a two-stage experiment, the first stage yielding J randomly sampled persons and the second stage consisting of the administration of I fixed test items (the test) to each sampled person. In this manner, dichotomous random variables $\{U_{ij}\}$; $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$ are generated. The basic statistical assumptions made are as follows:

1. Experimental independence of persons. The appropriate assumptions are made concerning the joint distribution of the $\{U_{ij}\}$ that correspond to the psychometric assumption that persons are randomly sampled from a very large population and that sampled persons respond to items independently of one another.
2. Local independence of items. The appropriate probabilistic assumptions are made concerning the joint distribution of the $\{U_{ij}\}$ and $\underset{\sim}{\theta}$ that correspond to the psychometric assumption that for each person, his or her responses to different items are independent.

Consider again the example of the introductory paragraph, that of a "mathematics" test. It might be that while θ is assumed to be a one-dimensional random variable measuring mathematical ability, in reality θ is two dimensional with the first dimension being mathematical ability and the second dimension being verbal ability. In the case of psychological testing, the most important statistical problem concerning dimensionality is to test $H : d = 1$ vs. $A : d > 1$. Recently, this author has constructed a statistic to test this hypothesis and to be further used as an index that estimates the amount of regularity in the data attributable to the multidimensionality of $\underset{\sim}{\theta}$.

Illustration

It is rather easy to imagine applications in other fields. As an illustration, suppose that medical subjects (the units) undergo allergy sensitivity tests to various environmental substances (each such test is a treatment). Suppose that the result of each test is scored 1 or 0, depending on whether an allergic reaction is observed or not. Let different values of the parameter θ be assigned to subjects according to each subject's sensitivity. Then, inferences about the dimensionality of Θ become meaningful in attempting to develop a classification scheme for allergies.

Description of the Statistic

A description of the constructed statistic can now be given. In doing so, the psychological testing language of items, persons, and so forth, will be used.

1. The test being administered is split into two subtests of lengths M and n, respectively. Here, n should be considered as large and M as possibly not large. Let f_n denote the proportion correct on the second subtest of items M + 1, M + 2, ..., M + n.
2. [0,1) is partitioned into intervals

   $$\bigcup_k A_n^{(k)} = [0,1) \qquad [8]$$

   such that

   $$\max_k \{\text{width } (A_n^{(k)})\} \to 0 \text{ as } n \to \infty . \qquad [9]$$

   For example, let

   $$A_n^{(k)} = \left[\frac{k-1}{[n^{1/2}]}, \frac{k}{[n^{1/2}]}\right] \qquad k = 1,2, \ldots, [n^{1/2}] \qquad [10]$$

   where [x] denotes the integer m such that $m \leq x$.
3. Persons are now grouped into categories according to the following rule: Assign a person to category (k,n) if for that person

   $$f_n \in A_n^{(k)} \qquad k = 1,2, \ldots, K_n . \qquad [11]$$

   (Here, K_n denotes the number of categories.) Thus, persons are assigned to the same category if they all get about the same proportion correct. This categorization of persons is the only use made of the second subtest. Let $J_n^{(k)}$ denote the number of persons in category (k,n).
4. To construct the test statistic, take the ratio of two variance estimators, the denominator estimating a variance that is uninfluenced by the "amount" of multidimensionality present and the numerator estimating a variance that is inflated by the amount of multidimensionality present. The variance estimators are each based upon the first subtest, i.e., on Items 1, 2, ..., M.
5. Now, fix (k,n). That is, look at the persons in cell k of the

_n_th partition $\{A_n^{(1)}, A_n^{(2)}, \ldots, A_n^{(K_n)}\}$, K_n denoting the number of partition cells.

6. The denominator can now be constructed. Consider item _m_ (of the first subtest, hence, $1 \leq m \leq M$. Let

$$\hat{P}_m^{(k)} = \sum_{j=1}^{J_n^{(k)}} U_{mj}/J_n^{(k)}, \qquad [12]$$

where U_{mj} indicates that correctness of the response of the _j_th person of cell _k_ to item _m_. Let

$$\hat{\sigma}_{Pk}^2 = \sum_{m=1}^{M} \hat{P}_m^{(k)}(1 - \hat{P}_m^{(k)})/M^2 , \qquad [13]$$

the denominator estimator of variance. Note that persons have been summed over first, forming $\hat{P}_m^{(k)}$ and _then_ items, forming $\hat{\sigma}_{Pk}^2$.

7. For the numerator, let $g_j^{(k)}$ be the proportion correct for person _j_ on the first subtest, i.e.,

$$g_j^{(k)} = \sum_{m=1}^{M} U_{mj}/M . \qquad [14]$$

Let

$$\bar{g}^{(k)} = \sum_{j=1}^{J_n^{(k)}} g_j^{(k)}/J_n^{(k)} \qquad [15]$$

and

$$\hat{\sigma}_{gk}^2 = \sum_{j=1}^{J_n^{(k)}} (g_j^{(k)} - {\bar{g}^{(k)}}^2)/J_n^{(k)} , \qquad [16]$$

the numerator estimator of variance. Note that items have been summed over first, forming $g_j^{(k)}$ and _then_ persons, forming $\hat{\sigma}_{gk}^2$.

8. For the estimator let

$$F_k = \hat{\sigma}_{gk}^2/\hat{\sigma}_{Pk}^2 . \qquad [17]$$

Thus, for each cell k, a statistic F_k is obtained. The $\{F_k\}$ are independent random variables.

The Asymptotic Distribution of $\{F_k\}$

In order to use the $\{F_k\}$ to make inferences about dimensionality, their asymptotic distribution is needed. To this end, the author has shown that for any K cells indexed by 1, 2, ..., K there exists $c_k > 0$ such that

$$\sum_{k=1}^{K} \frac{F_k - 1}{c_k} \Big/ \sqrt{K} \qquad [18]$$

is asymptotically normal with mean zero and variance one [notationally N(0,1)] when d = 1 and, moreover, estimators $\hat{c}_k$ of c_k exist such that

$$\sum_{k=1}^{K} \frac{F_k - 1}{\hat{c}_k} \Big/ \sqrt{K} \qquad [19]$$

is asymptotically N(0,1) when d = 1. Further, it has been shown that there exists a number $C > 0$ and numbers $A_{M,k} \geq CM$ such that $F_k \to A_{M,k}$ in probability for k = 1, 2, ..., K when $d > 1$. Hence, there exists a valid large sample level α procedure for testing $H : d = 1$ vs. $A : d > 1$.

It also follows that this procedure (even in the extreme case of K = 1) for an appropriate choice of M has asymptotic power one for any fixed alternative, i.e., any distribution of $\underset{\sim}{\theta}$ for which $d > 1$. The procedure is to reject H if

$$\sum_{k=1}^{K} \frac{F_k - 1}{\hat{c}_k} \Big/ \sqrt{K} > Z_\alpha , \qquad [20]$$

where Z_α is the 100 $(1 - \alpha)$ percentile of a standard normal distribution.

Discussion

There remain several important theoretical and practical questions that should be investigated. First, there are clearly several plausible ways of combining the F_k's into a single test statistic and of obtaining the asymptotic distribution of this test statistic. Three such possibilities are

1. $\Sigma_k \frac{F_k - 1}{\hat{c}_k} \Big/ \sqrt{K}$

   as was shown above;

2. $\Sigma_k\left(I\left[\frac{F_k-1}{\hat{c}_k} > Z_\alpha\right] - \alpha\right)\Big/\sqrt{K}$

   where I[A] denotes the indicator of the event A; and

3. A chi-square like statistic $[\Sigma_k(O_k - F_k)^2/E_k]$ based upon the number of k's such that

   $$\frac{F_k-1}{\hat{c}_k} > Z_\alpha .$$

The author plans to investigate the asymptotic distributions of the second and third of these statistics as well.

Second, it is essential to carry out some carefully designed monte carlo studies to see for what range of test lengths and sample sizes of examinees the actual distribution of the F_k's is well approximated by the asymptotic distribution of the F_k's. This is essential because asymptotic distribution theory cannot by itself guarantee the accuracy of the approximation that it suggests.

Third, the meaningful and practical question is not whether d = 1 but, rather, whether taking d = 1 accounts for most of the explainable regularity in the data. Thus, what is called for is a reformulation of the hypothesis that d = 1 and possibly an estimation approach in order to estimate how much of the explainable regularity is accounted for by taking d = 1. This important practical concern needs to be dealt with by some combination of a theoretical analysis and a monte carlo study.

Fourth, some combination of a theoretical analysis and a monte carlo study is also needed so that some quantitative information is available about the power of the tests constructed from the F_k's.

Fifth, the "regularity" conditions that were needed on the rate of growth of the $\{J_n^{(k)}\}$ (numbers of persons per cell) as $n \to \infty$ in order to establish the asymptotic normality of the F_k's--and, hence, the asymptotic distribution of the statistics described above--can undoubtedly be improved upon. This would further strengthen the case for using the F_k's in actual testing situations. Moreover, it is quite possible that the methods of proof used or the results obtained when abstracted from the present situation involving the F_k's may add to the general body of knowledge in mathematical statistics.

Sixth, the procedures that are obtained from carrying out the above should be pilot tested on actual tests and populations.

Seventh, a thorough comparison between these procedures based on the F_k's and on any other approaches (such as factor analytic) in the literature must be made.

Finally, procedures should also be developed for testing $H : d = k$ vs. $A : d > k$ for fixed $k \geq 2$. Although the derivation of the distribution of the F_k's under the assumption $d = 1$ was surprisingly delicate, it seems clear that an analogous procedure for this hypothesis testing situation can be found and its properties studied.

The author plans to investigate these questions with the goal of producing a theoretically sound and practically important statistical approach to the problem of making inferences about the underlying dimensionality.

REFERENCES

Levine, M. V. Item-item curves and consistent mental test parameter estimates (ETS RB-76-36). Princeton NJ: Educational Testing Service, 1976.

Levine, M. V. Personal communication, 1981.

ACKNOWLEDGMENTS

This work was partially supported by the Office of Naval Research (N00014-79-C-0752; NR 150-445) and by the National Science Foundation.

APPLICATION OF UNIDIMENSIONAL ITEM RESPONSE THEORY MODELS TO MULTIDIMENSIONAL DATA

FRITZ DRASGOW
UNIVERSITY OF ILLINOIS

CHARLES K. PARSONS
GEORGIA INSTITUTE OF TECHNOLOGY

Lord and Novick (1968) have stated, "it can be taken for granted that every [item response theory] model is false and that we can prove it so, if only we collect a sufficiently large sample of data" (p. 383). One way in which currently available item response theories are surely incorrect is in their assumption of a unidimensional latent trait space. This assumption, which implies local independence of item scores, is an essential part of the theory underlying most currently available parameter estimation procedures.

The implications of violations of the unidimensionality assumption for testing practitioners and substantive researchers require careful examination. The present research examines the effects of a multidimensional latent trait space on estimation of item and person parameters by the widely used computer program LOGIST (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976). LOGIST uses the method of maximum likelihood and a number of ad hoc techniques to estimate item and person parameters of the 1-, 2-, or 3-parameter logistic models. The assumption of unidimensionality is used in the derivation of the likelihood function maximized by LOGIST (see Lord, 1980, p. 19). Consequently, estimates computed by LOGIST have a theoretical justification only in the case of a unidimensional latent trait space.

From a practical perspective it is important to know the extent to which LOGIST's parameter estimates are robust to violations of unidimensionality. In particular, when is an item pool "sufficiently unidimensional" for parameter estimates to be useful to testing practitioners and substantive researchers?

The formalization of the notion of "sufficiently unidimensional" used in the present research was based on three examples:

1. A test of verbal ability that has subsections composed of antonyms, analogies, and paragraph comprehension questions;
2. A test of algebra achievement that asks questions based on each of several parts of a high school algebra course; and
3. An instrument measuring overall job satisfaction that asks questions

about workers' affective responses to a number of job characteristics including their supervision, pay, and coworkers.

In each of these examples, interest is centered upon a single latent trait that underlies responses to all items in the item pool. However, it is clear that none of the three instruments is truly unidimensional. In particular, clusters of items are likely to be more highly related than expected on the basis of a single latent trait. An item pool is defined to be "sufficiently unidimensional" to allow application of an item response theory (IRT) model and estimation procedure if the estimation procedure recovers the general latent trait underlying responses to all items in the item pool.

Previous research studying the accuracy of LOGIST has generally used simulated item responses that meet the assumptions (including unidimensionality) of the model fitted to the data. Under these conditions, Lord (1975), Swaminathan and Gifford (1979), Hulin, Lissak, and Drasgow (1982) and others have found that LOGIST provides effective parameter estimation when sample size (N) and test length (n) are sufficiently large. Provided that the assumptions are satisfied, $N \geq 1000$ and $n \geq 50$ appear to be adequate for the 3-parameter logistic model. Minimum requirements for the 2-parameter logistic model are less restrictive.

Reckase (1979) conducted one of the few studies that examined the effectiveness of LOGIST (or any estimation technique) when the unidimensionality assumption is violated. He generated a data set with an underlying dominant latent trait that was related to all items, as well as weaker latent traits that affected clusters of items. LOGIST was found to be robust to these minor violations of the unidimensionality assumption in the sense that the dominant latent trait was well recovered. Reckase also simulated a test composed of items that were factorially pure measures of two statistically independent latent traits. Here LOGIST was drawn to one of the two latent traits: LOGIST's ability estimates were highly correlated ($r = .93$) with estimated factor scores on one factor and nearly uncorrelated ($r = .29$) with estimated factor scores on the other factor. In addition, LOGIST's estimates of the item discrimination parameter were generally greater than 1.70 for items related to the former trait and less than .15 for items related to the latter trait.

It is useful to think of the prepotency of a general latent trait as varying along a continuum. At one extreme the latent space is truly unidimensional. Reckase's case of a dominant general trait corresponds to a small move along the prepotency continuum away from unidimensionality. Note that this simulated item pool was sufficiently unidimensional for LOGIST to recover the general trait. Reckase's hypothetical item pool composed of factorially pure items measuring two independent traits lies at the other extreme of the continuum. Here there was no general factor, which is equivalent to a general factor with zero prepotency. It is obviously impossible for LOGIST (or any estimation technique) to recover a general trait at this end of the prepotency continuum.

In the present research, several item pools were simulated that ranged from truly unidimensional to an inconsequential general latent trait. Item pools with intermediate levels of prepotency of the general latent trait were also constructed. These item pools were used to determine the degree of prepotency that is required by LOGIST in order to recover the general latent trait and not be drawn to a latent trait underlying a cluster of items.

Method

Simulation Model and Parameters

The simulation model used in the present research consisted of the following components. First, correlated common factors were simulated using the hierarchical factor analysis model proposed by Schmid and Leiman (1957). A single second-order *general* factor controlled the correlations of the first-order common factors. Latent *item propensity scores* were generated for *n* hypothetical items using the underlying factors. Finally, dichotomous item scores were created by determining whether item propensity scores were above or below their respective threshold values.

The factor model. The common factor model can be written

$$\underset{\sim}{x} = \underset{\sim}{A}\underset{\sim}{y} + \underset{\sim}{B}\underset{\sim}{e} \qquad [1]$$

where

$\underset{\sim}{x}$ is a vector containing the *n* observed variables x_i (here, however, "observed" variables were not observed but, instead, were considered as item propensity variables that underlie observed item responses);

$\underset{\sim}{A}$ is the $n \times k$ matrix of loadings on the *k* common factors;

$\underset{\sim}{y}$ is a vector containing the *k* common factors y_i;

$\underset{\sim}{B}$ is an $n \times n$ diagonal matrix with loadings on unique factors along its diagonal; and

$\underset{\sim}{e}$ is a vector containing the *n* unique variables e_i.

The unique variables are assumed to be mutually uncorrelated and uncorrelated with the common factors. Let α_{ij} denote the loadings of the *i*th item propensity variable on the *j*th factor and let β_i denote the single loading of the *i*th item propensity variable on the *i*th unique variable. The item propensity variables, common factors, and unique factors were all scaled to have mean zero and unit variance.

The factor loading matrix $\underset{\sim}{A}$ used in the present research is shown in Table 1. Note that each of the 50 item propensity variables loads on a single common factor and that there are five common factors. The first common factor is related to 15 item propensity variables, the second is related to 5 item propensity variables, and the rest are related to 10 item propensity variables. Since the magnitudes of the factor loadings are comparable across all five common factors, it is apparent that the first common factor is the most influential in this particular item pool.

The simple pattern of factor loadings in Table 1 was selected for several interrelated reasons. First, simple structure (Thurstone, 1947, chap. 14) is a convenient means for specifying a frame of reference in the factor space that eliminates the rotational indeterminacy of factors. Moreover, rotations to approximate simple structure seem possible for each of the three examples de-

Table 1
Factor Loading Matrix A and Item Threshold Values (γ)

Item Propensity Variable	Common Factor					γ	Item Propensity Variable	Common Factor					γ
	1	2	3	4	5			1	2	3	4	5	
1	.4					.85	26			.4			-.53
2	.5					-.53	27			.5			1.30
3	.6					-.85	28			.6			-.85
4	.7					-.26	29			.7			-.26
5	.8					.26	30			.8			.26
6	.4					-1.30	31				.4		-1.30
7	.5					.53	32				.5		.00
8	.6					-.85	33				.6		.85
9	.7					.00	34				.7		.53
10	.8					1.30	35				.8		-1.30
11	.4					1.30	36				.4		-.85
12	.5					-1.30	37				.5		.26
13	.6					.53	38				.6		-.26
14	.7					-.53	39				.7		-.53
15	.8					.00	40				.8		1.30
16		.4				-.85	41					.4	.53
17		.5				-1.30	42					.5	1.30
18		.6				.85	43					.6	-1.30
19		.7				.00	44					.7	-1.30
20		.8				-.26	45					.8	-.26
21			.4			-1.30	46					.4	-.53
22			.5			-1.30	47					.5	.85
23			.6			.53	48					.6	.26
24			.7			.85	49					.7	-.85
25			.8			.00	50					.8	.00

Note. Only nonzero factor loadings are shown; zero loadings have been left as blanks.

scribed previously. With actual data from ability tests, achievement tests, and attitude assessments, it is usually necessary for the common factors to be correlated in order to rotate to simple structure. Let Φ denote the matrix of correlations of first-order common factors after rotation to oblique simple structure.

A fundamental assumption of the simulation model is that a single second-order general factor accounts for the first-order common factor correlations in Φ. The substantive motivation for this assumption can be seen by again considering the examples previously mentioned. For each of the three instruments, all items should be affected by a general latent trait: general verbal ability, general algebra achievement, and overall job satisfaction, respectively. The way in which the IRT unidimensionality assumption is violated is by clusters of items with larger within-cluster correlations than would be expected on the basis of a single first-order common factor. To account for these within-cluster correlations, it is usually necessary to extract several first-order common factors. Rotation to approximate simple structure might yield a factor loading matrix analogous to the idealized A matrix in Table 1 and a factor correlation matrix Φ with large factor intercorrelations. If Φ were then factor analyzed, a single general factor (i.e., a second-order common factor) would emerge. This would result because all items are affected by a single general latent trait.

Schmid and Leiman (1957) proposed a convenient model for the factor analysis of correlated common factors. A special case of this general model was used in the present research. Here, common factors y are represented by

$$y = fz + Dv \qquad [2]$$

where

- f is the k element vector containing loadings of the first-order common factors on the single general factor z,
- D is a $k \times k$ diagonal matrix containing loadings of the first-order common factors on the k second-order group factors (i.e.; second-order specific factors; and
- v is a vector containing the k group factors.

The group factors are assumed to be mutually uncorrelated and uncorrelated with z. Again, all factors are scaled to have unit variance.

The second-order general factor z can be interpreted as the general latent trait underlying responses to all items in the item pool. The elements v_i of v correspond to additional latent variables, uncorrelated with z, that cause clusters of items to be more highly related than would be expected on the basis of a unidimensional latent trait. For example, if z were general verbal ability, then v_j might represent an individual's ability to solve, say, analogies, after controlling for general verbal ability.

The elements f_j of f and diagonal elements δ_j of D control the relative importance of the general and group factors. The combinations of f and D used in the present research are shown in Table 2. The first combination shown in Table 2, labeled Latent Structure 1, has five common factors that are perfectly correlated; the latent trait space is therefore truly unidimensional. Results for this latent structure were used as baseline values; it is unlikely that actual data sets would ever be truly unidimensional. The prepotency of the general factor gradually decreases across the remaining four combinations. Latent Structure 2 was designed to simulate the test of verbal ability described previously. The oblique common factors, which might correspond to factors associated with the various item types, have intercorrelations that range from .68 to .90. Analytic rotation methods designed to rotate to oblique simple structure would be likely to encounter difficulties in recovering such highly correlated factors. Moderately heterogeneous achievement tests and attitude assessment instruments are simulated by the third latent structure. Here intercorrelations of the oblique common factors range from .46 to .60. Latent Structure 4 was designed to simulate broad range achievement tests and attitude assessment instruments. The oblique common factors from this latent structure have intercorrelations that range from .25 to .39. Finally, the general factor from Latent

Structure 5 is very weak, perhaps corresponding to "method variance," rather than to a psychologically meaningful trait.

Table 2
Loadings of Common Factors on General and Group Factors

Common Factor	Latent Structure 1		2		3		4		5	
	f_j	δ_j	f_j	δ_j	f_j	δ_j	f_j	δ_j	f_j	δ_j
1	1.00	.00	.95	.31	.70	.71	.65	.76	.40	.92
2	1.00	.00	.90	.44	.65	.76	.55	.84	.35	.94
3	1.00	.00	.85	.53	.70	.71	.60	.80	.10	.99
4	1.00	.00	.80	.60	.80	.60	.45	.89	.20	.98
5	1.00	.00	.95	.61	.75	.66	.55	.84	.25	.97

Note. Correlations between common factors ranged from .68 to .90 for Latent Structure 2, from .46 to .60 for Latent Structure 3, from .25 to .39 for Latent Structure 4, and from .02 to .14 for Latent Structure 5.

Response generation. The item propensity variables x_i in Equation 1 can be obtained directly from the second-order general and group factors and first-order unique factors by

$$\begin{aligned}\underset{\sim}{x} &= \underset{\sim}{A}\,[\underset{\sim}{f} \vdots \underset{\sim}{D}]\begin{bmatrix} z \\ \cdots \\ \underset{\sim}{v} \end{bmatrix} + \underset{\sim\sim}{Be} \\ &= [\underset{\sim}{g} \vdots \underset{\sim}{S}]\begin{bmatrix} z \\ \cdots \\ \underset{\sim}{v} \end{bmatrix} + \underset{\sim\sim}{Be} \qquad [3]\end{aligned}$$

where $\underset{\sim}{g} = \underset{\sim\sim}{Af}$ is the n element vector containing loadings of item propensity variables on the general factor and $\underset{\sim}{S} = \underset{\sim\sim}{AD}$ is the $n \times k$ matrix containing loadings of item propensity variables on the group factors. When $\underset{\sim}{A}$ has a simple pattern of loadings as in Table 1, the ith diagonal element of $\underset{\sim}{B}$ can be computed by

$$\beta_i = \sqrt{1 - \Sigma_j \alpha_{ij}^{\;2}} \; . \qquad [4]$$

In the present research Equation 3 was used to generate the item propensity variables and $\underset{\sim}{A}$, $\underset{\sim}{f}$, and $\underset{\sim}{D}$, from Tables 1 and 2 were used to compute $\underset{\sim}{g}$ and $\underset{\sim}{S}$. Note that the first group factor v_1 is related to 15 item propensity variables, v_2 is related to 5 item propensity variables, and v_3, v_4, and v_5 are each related to 10 item propensity variables. The group factor z, the v_i, and the e_i were all generated as independent normal (0,1) variables by the IMSL (1975) subroutine GGNPM.

Item responses u_i were simulated by dichotomizing the item propensity variables:

$$u_i = \begin{cases} 1 \text{ if } x_i > \gamma_i \\ 0 \text{ if } x_i < \gamma_i \ . \end{cases} \qquad [5]$$

The threshold values γ_i are presented in Table 1. These values were sampled from a uniform distribution on approximately the nine decile points of the normal distribution.

Data Sets

Samples of N = 1,000 simulated examinees were created using Equations 3 and 5 for each of the five latent structures in Table 2. These samples are labeled Data Sets 1 through 5, respectively. Note that item responses were generated by a process that is exactly equivalent to the 2-parameter normal ogive model (Lord & Novick, 1968, chap. 16) when the truly unidimensional Latent Structure 1 is used in Equation 3, i.e., in Data Set 1.

Guessing can be simulated by first generating item responses by Equations 3 and 5 and then, if $u_i = 0$, by rescoring u_i to be correct with probability c_i. Item responses with guessing for samples of N = 1,500 were generated using each of the five latent structures in Table 2, with $c_i = .15$ for even-numbered items and $c_i = .20$ for odd-numbered items. These samples are labeled Data Sets 6 through 10.

Criteria for Evaluation of Parameter Estimates

Item parameters. Lord and Novick (1968, p. 375) derived two important relations between the 2-parameter normal ogive model and the factor analysis model. For Data Set 1 the methods used in the present simulation corresponded exactly to the assumptions made by Lord and Novick. Using these assumptions Lord and Novick showed that

$$a_i = \frac{g_i}{\sqrt{1 - g_i^2}} \qquad [6]$$

where a_i is the item discrimination parameter for item _i_, and g_i is the loading of the _i_th item propensity variable x_i on the general factor. They also proved that the item difficulty parameter b_i is

$$b_i = \frac{\gamma_i}{g_i} \ . \qquad [7]$$

Note that Equation 7 implies that b_i is undefined when g_i is zero.

BEST COPY AVAILABLE

For the multidimensional Data Sets 2 through 5, Equation 6 can be applied to the loading g_i on the general factor. An equation analogous to Equation 6 can also be defined as

$$\tilde{a}_{ij} = \frac{s_{ij}}{\sqrt{1 - s_{ij}^2}} \qquad [8]$$

where s_{ij} is the loading of the *i*th item propensity variable x_i on the *j*th group factor v_j in Equation 3.

It appears reasonable to conclude that LOGIST is robust to violations of unidimensionality to the extent that the elements of *g*, after transformation by Equations 6 and 7, are related to estimates of *a* and *b* and that the elements of S, after transformation by Equation 8, are *not* related to estimates of *a*. The measures of association used in the present research were root mean squared differences (RMSDs)

$$\text{RMSD for } a = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{a}_i - a_i)^2}\ , \qquad [9]$$

$$\text{RMSD for } \tilde{a}_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{a}_i - \tilde{a}_{ij})^2}\ , \qquad [10]$$

and

$$\text{RMSD for } b = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{b}_i - b_i)^2} \qquad [11]$$

where $\hat{a}_i$ and $\hat{b}_i$ are estimates of a_i and b_i, respectively, obtained from LOGIST. Equations 6 and 7 can also be used when there is guessing. Thus, the RMSDs in Equations 9 through 11 can be computed, and the RMSDs for Data Sets 1 and 6 can be used as baseline values for interpreting RMSDs for Data Sets 2 through 5 and 7 through 10.

Person Parameters

Lord and Novick (1968) showed that the person parameter θ of the 2-parameter normal ogive model is the general factor *z* under conditions equivalent to those implied by the first latent structure of Table 2 when there is no guessing. Their results can also be used to derive the same relation when guessing occurs. Again, it is reasonable to conclude that LOGIST is robust to violations of unidimensionality to the extent that its estimates $\hat{\theta}$ of person parameters are strongly related to *z* and are not related to the group factors v_i. Convenient measures of association are the product-moment correlations $r(\hat{\theta}, z)$ between $\hat{\theta}$ and *z* and $r(\hat{\theta}, v_i)$ between $\hat{\theta}$ and each of the v_i.

Results

Item Parameters

Estimates of a and b parameters were obtained for Data Sets 1 through 5 (i.e., the data sets without guessing) by LOGIST using default convergence criteria. Values of c_i were fixed at 0 so that LOGIST was fitting the 2-parameter logistic model to item responses. (Although LOGIST used the logistic model, Birnbaum [1968, p. 399] showed that appropriately scaled logistic models are virtually undistinguishable from the normal ogive model used to generate item responses.)

Based on earlier studies (e.g., Hulin et al., 1982), it was expected that the numbers of items (n = 50) and examinees (N = 1,000) would be large enough for accurate estimation of item parameters when the unidimensionality assumption was satisfied. Table 3, which presents item parameter RMSDs for Data Sets 1 through 5, shows that N and n were indeed sufficiently large. For Data Set 1 the RMSD for a is .136 and the RMSD for b is .220, both of which are small enough for most practical purposes.

Table 3
Root Mean Squared Differences Obtained by
Equations 9, 10, and 11 for Responses with No Guessing

Factor	Data Set 1 $\hat{a}$	1 $\hat{b}$	2 $\hat{a}$	2 $\hat{b}$	3 $\hat{a}$	3 $\hat{b}$	4 $\hat{a}$	4 $\hat{b}$	5 $\hat{a}$	5 $\hat{b}$
z	.136	.220	.149	.292	.172	.474	.223	.755	.455	4.093
v_1			.775		.519		.377		.225	.324*
v_2			.811		.638		.569		.628	
v_3			.778		.603		.544		.696	
v_4			.782		.598		.561		.678	
v_5			.796		.594		.538		.668	

*RMSD for b computed only for Items 1 to 15. Values of s_{i1} were used in Equation 7 in place of g_i.

The effects of the decreasing importance of the general factor in determining item responses across Latent Structures 2 through 5 are clear in Table 3. The RMSD of a increases only a small amount from the baseline value of .136 for Data Set 1 to .172 for Data Set 3. Then there is a moderate increase to .223 in Data Set 4 and a large increase to .455 in Data Set 5. The first group factor v_1 which underlies responses to the largest number (15) of items, has a RMSD for $\hat{a}_1$ that is markedly lower than any of the other group factors in Data Set 4. In Data Set 5, RMSD for $\hat{a}_1$ is much lower than the RMSD for the general factor. Thus, in Data Set 4 it appears that LOGIST has been drawn to a latent trait that is a composite of the general factor and the most influential group factor. It is clear that LOGIST has been drawn to the first group factor in the fifth data set.

These conclusions are further supported by examining the estimates of *b*. In Data Set 5, values of $\hat{b}$ seem to be approaching the values that would result from applying Equation 7 to the loadings of items on the first group factor. Values of $\hat{b}$ for the 15 items with nonzero loadings on the first group factor are quite close to the factor loadings on the first group factor transformed by Equation 7: The RMSD of *b* for these 15 items was .324. In contrast, values of $\hat{b}$ (not shown) were excessively large for the 35 items, with zero loadings on the first group factor. Application of Equation 7 to an item with a zero loading produced an infinite value of *b*. Thus, the large values of $\hat{b}$ obtained for items with zero loadings on the group factor support the conclusion that LOGIST has been drawn to the first group factor in Data Set 5.

The results for the five data sets with guessing are shown in Table 4. Default convergence criteria were again used for LOGIST, but $\hat{c}_i$ was free to vary as would be appropriate if LOGIST were used to estimate parameters of multiple-choice test items. Although the RMSDs for *a* increase more quickly in Table 4 than in Table 3, the pattern of results is generally similar. In Data Set 8 the RMSD for *a* (.316) is much smaller than the RMSD for $\tilde{a}_1$(.648). These two RMSDs are much more similar in magnitude in Data Set 9: .402 and .491. In Data Set 10 the RMSD for $\tilde{a}_1$ is substantially smaller than the RMSD for *a*.

Table 4
Root Mean Squared Differences Obtained by
Equations 9, 10, and 11 for Responses with Guessing

	Data Set									
	6		7		8		9		10	
Factor	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
z	.209	.194	.247	.368	.316	.735	.402	.817	.535	9.374
v_1			.847		.648		.491		.319	
v_2			.890		.758		.702		.704	
v_3			.864		.721		.688		.761	
v_4			.867		.697		.720		.743	
v_5			.873		.699		.698		.721	

The RMSD for *b* in Data Set 8 is .735, a value that is considerably larger than the corresponding RMSDs for Data Sets 6 and 7. This large value results in part from two sources. First, it is evident that as loadings on the general factor decrease, values of b_i in Equation 7 increase. For example, Item 11 has b_{11} values of 3.25, 3.42, and 4.64 in Data Sets 6, 7, and 8. Second, Hulin et al. (1982) found that it is very difficult to estimate parameters of 3-parameter logistic items with values of *b* that are large in magnitude. This finding is not particularly surprising. An unexpected result obtained by Hulin et al., however, is that the problems encountered with 2-parameter logistic item responses are much less severe. For example, with a simulated test of 1,000 examinees and 30 items, Hulin et al. (using exactly the same population values of *a* and *b* when simulating 2- and 3-parameter logistic item responses) found that the correlation between estimated and actual b_i values was .995 for 2-parameter logistic item responses but only .623 for 3-parameter logistic item responses. Interestingly, the correlation between estimated and actual b_i values was .949 for 3-parameter logistic items with $|b_i| < 2.0$. Thus, it is extreme values of b_i that are particularly difficult to estimate for the 3-parameter logistic model. In analysis of variance terminology, there is an interaction between item response model and magnitude of b_i in determining the accuracy of estimation of b_i.

In sum, the large Data Set 8 RMSD for $\underline{b}$ could be due to the interaction of item response model and b_i magnitude (which is relevant because multidimensionality has caused many b_i values to become large in magnitude) or due to LOGIST being drawn to the first group factor. From the results concerning the RMSD for $\underline{a}$, it appears that the former interpretation should be adopted.

Person Parameters

The correlations between estimates $\hat{\theta}$ of ability computed by LOGIST and factor scores appear in Table 5. These correlations were computed from the actual factor scores used in Equation 3—not factor score estimates.

Table 5
Correlations Between Ability Estimates
and Factor Scores

Data Set	Factor					
	z	v_1	v_2	v_3	v_4	v_5
1	.965					
2	.939	.044	.071	.173	.076	.053
3	.843	.322	.078	.154	.129	.153
4	.736	.461	.061	.139	.212	.225
5	.376	.774	.064	.039	.062	.083
6	.939					
7	.909	.119	.029	.078	.097	.096
8	.828	.278	.044	.118	.136	.206
9	.722	.444	.126	.145	.085	.127
10	.348	.709	−.030	.015	.056	.153

Note. Correlations between group factors and ability estimates were zero for Data Sets 1 and 6.

The results seen in Table 5 are orderly and compelling: As the prepotency of the general factor decreases, $r(\hat{\theta},z)$ values decrease and $r(\hat{\theta},v_1)$ values increase. For both data sets with guessing and data sets without guessing, $\hat{\theta}$ is virtually unrelated to v_1 when item responses are generated from the second latent structure shown in Table 2. The correlations between v_1 and $\hat{\theta}$ are only about .3 when item responses are generated from the third latent structure.

Here $r(\hat{\theta},z)$ is about .83, which is large enough for many practical applications. In Data Sets 4 and 9, $\hat{\theta}$ values are more highly related to the general factor than the first group factor, but it is clear that LOGIST's ability estimates are strongly influenced by both $\underline{z}$ and v_1. It is apparent that LOGIST has been drawn to v_1 in Data Sets 5 and 10.

Discussion

The types of multidimensionality studied here have several effects on the estimation techniques programmed in LOGIST. Perhaps most important is that as the prepotency of the general factor decreases, LOGIST is gradually drawn to the strongest group factor. RMSDs for $\underline{a}$ and RMSDs for $\underline{b}$ increase, slowly at first and then more rapidly, as the latent structure varies across Levels 1 through 5 in Table 2. As the prepotency of the general factor decreases, the effects on the RMSDs are paralleled by decreasing correlations between $\hat{\theta}$ and $\underline{z}$ and increasing correlations between $\hat{\theta}$ and v_1.

Estimates of item difficulty occasionally become excessively large in magnitude when actual data sets are analyzed by LOGIST although the most recent version of LOGIST [Wingersky, in press] has options that may reduce this problem). The results obtained here indicate that this phenomenon may partially be due to multidimensional item pools. Of course, these items may be poorly written, too easy, or too difficult. Nonetheless, Equation 7 and Tables 3 and 4 indicate that decreasing the prepotency of the general factor (for example, by increasing the number of content areas of an achievement test) may cause some items to have values of $\underline{b}$ that are very large in magnitude.

Latent Structure 2 of Table 2 was originally designed to simulate a very homogeneous test such as a test of verbal ability that uses several types of items. From Tables 3, 4, and 5 it is clear that LOGIST is robust to the minor violations of multidimensionality seen in Data Sets 2 and 6.

The third latent structure in Table 3 was designed to simulate more heterogeneous measurement, such as an instrument measuring overall job satisfaction or algebra achievement. Here about 70% of the variance in $\hat{\theta}$ is due to the general factor and only about 10% is due to the strongest group factor. Moreover, comparing RMSDs of $\underline{a}$ for data sets based on the third latent structure to baseline RMSDs of $\underline{a}$ shows that LOGIST still recovers the item discrimination parameter implied by the general factor. Consequently, it appears reasonable to conclude that use of LOGIST is justified in item pools with multidimensionality of the type seen in Data Sets 3 and 8. The common factors in Equation 1 that underlie Data Sets 3 and 8 item responses have correlations from .46 to .60. This means that these simulated item pools are quite heterogeneous. Note that factor analyzing the dichotomous item responses of Data Sets 3 and 8 may yield factors with intercorrelations smaller than .46 to .60 and that guessing may further decrease factor correlations.

Use of LOGIST in Data Sets 4 and 9 leads to parameter estimates with interpretations that are ambiguous at best. In Data Sets 5 and 10, LOGIST is clearly drawn to a group factor. In sum, it appears that LOGIST should not be used in data sets with the degree of multidimensionality seen in the fourth and fifth latent structures of Table 2.

The results obtained here indicate that retaining the null hypothesis that an item pool is unidimensional when conducting a significance test, such as the ones developed by Christoffersson (1975) and Muthén (1978), is not a prerequisite for applications of IRT. A powerful significance test would always reject the null hypothesis of unidimensionality for large samples of actual examinees. Nonetheless, results for Data Sets 2, 3, 7, and 8 indicate that LOGIST parameter estimates will have many practically useful applications in multidimensional item pools.

One criticism of the research described in this paper is that unidimensional IRT models are improperly applied to multidimensional item pools; instead, it would be argued that multidimensional IRT models (see Reckase & McKinley, 1983; Sympson, 1978) should be used for multidimensional item pools. If a single dominant latent trait is not sufficiently prepotent, the results presented here clearly show that a unidimensional model is inadequate. However, it is important to note that unidimensional models do provide a good description of multidimensional data sets when the dominant latent trait is sufficiently prepotent. Moreover, it appears that robustness studies of the type described here will be necessary when workable estimation methods for multidimensional IRT models become available: A small number of interpretable latent traits will not span the entire latent trait space underlying item responses. This point is illustrated by the results of Christoffersson's (1975) significance tests (with $\alpha = .01$) for the number of common factors underlying a pool of 12 items. His significance tests showed that there were more than four common factors, of which only two factors were interpretable. Of course, more items could be written in an attempt to study these uninterpretable common factors; but it seems likely that further significance tests would then show that even more common factors were required and that some of these additional factors would be uninterpretable. Furthermore, the additional factors seem unlikely to provide valuable contributions to substantive theory. Thus, in the present context it seems clear that researchers should be more concerned with the robustness of estimation techniques to minor violations of dimensionality assumptions than with the possibly never-ending task of measuring all latent variables that underlie responses in a particular content domain.

It is important for researchers to investigate the dimensionality of an item pool before applying IRT. One obvious indication that multidimensionality is too severe for LOGIST is relatively many items with $\hat{b}$ values that are excessively large. A second, more sophisticated, approach for examining the latent structure of dichotomous item responses with or without guessing and with possibly nonnormal ability distributions is currently under investigation (Drasgow & Lissak, 1982).

It is important to note a number of limitations on the results obtained here. First, item responses were simulated using exactly one second-order general factor. The effects of two or more second-order general factors on parameter estimates computed by LOGIST are unknown. A multidimensional IRT model may be essential to adequately model data sets with two or more second-order general factors. Second, the magnitudes of the factor loadings presented in Table 1 are comparable across the five first-order common factors. Presumably, relatively smaller loadings on one common factor would reduce the influence of its underlying second-order group factor on parameter estimates; however, the details of the relations between LOGIST's parameter estimates and magnitudes of loadings on common factors are not clear. Third, only one pattern of factor loadings with 15, 5, 10, 10, and 10 items per factor was examined. The effects of a wider range than the 5 to 15 items per factor and of different distributions than 5, 10, 10, 10, and 15 require further investigation. A fourth limitation lies in the use of the idealized factor loading matrix shown in Table 1. Actual items are likely to have small but nonzero loadings on several factors and it is common to find that some items have large loadings on more than a single factor.

REFERENCES

Birnbaum, A. Some latent trait models. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Christoffersson, A. Factor analysis of dichotomized variables. Psychometrika, 1975, 40, 5-32.

Drasgow, F., & Lissak, R. I. Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Manuscript in preparation, 1982.

Hulin, C. L., Lissak, R. I., & Drasgow, F. Recovery of two and three parameter logistic item characteristic curves: A monte carlo study. Applied Psychological Measurement, 1982, 6, 249-260.

International Mathematical and Statistical Libraries. IMSL Library 1 (5th ed.). Houston TX: Author, 1975.

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (Research Bulletin 75-33). Princeton NJ: Educational Testing Service, 1975.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale NJ: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Muthén, B. Contributions to factor analysis of dichtomous variables. Psychometrika, 1978, 43, 551-560.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4, 207-230.

Reckase, M. D., & McKinley, R. L. Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1985.

Schmid, J., & Leiman, J. The development of hierarchical factor solutions. *Psychometrika*, 1957, *22*, 53-61.

Swaminathan, H., & Gifford, J. A. *Estimation of parameters in the three-parameter latent trait model* (Research Report No. 90). Amherst: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluation Research, 1979.

Sympson, J. B. A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Thurstone, L. L. *Multiple-factor analysis*. Chicago: University of Chicago Press, 1947.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *ERIBC monograph on application of item response theory*. Vancouver BC: Educational Research Institute of British Columbia, in press.

Wood, R. L., & Lord, F. M. *A user's guide to LOGIST* (Research Memorandum 76-4). Princeton NJ: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOGIST--A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76-6). Princeton NJ: Educational Testing Service, 1976.

ACKNOWLEDGMENTS

Thanks are due Charles L. Hulin for his comments on an earlier draft. The order of authorship was determined by a computer simulation of the toss of a fair coin.

Tools of Robustness for Item Response Theory

Douglas H. Jones
Educational Testing Service

The extension of robustness theory (Huber, 1981) to item response theory (IRT) holds great potential for testing. One primary benefit will be increased understanding of the behavior of the usual estimators of item and ability parameters when the assumed model does not hold. Another benefit will be new estimators resistant to departures from a chosen model.

The purpose of this paper is to briefly demonstrate just a few of the possibilities of a systematic application of robustness theory, concentrating on the estimation of ability when the true item response model does and does not fit the data. First, the definition of the maximum likelihood estimator (MLE) of ability will be briefly reviewed. Secondly, after introducing the notion of *outlier-pair* observation, the effect of the model's lack of fit to the data will be demonstrated. This will provide motivation for the introduction of a new estimator not as influenced as the MLE by outlier pairs. Based on the motivation for the new estimator, an entire class of new estimators, labeled *h* estimators, will then be introduced.

The *h* estimators will provide a vehicle for understanding two important notions of robustness theory: (1) the *influence function*, which compactly yields the majority of the behavioral results about an estimator and (2) *asymptotic relative efficiency* (ARE), a well-known concept in classical statistics and one which is of use in IRT.

Although the notion of ARE has had limited application in IRT, it is a useful alternative to expensive monte carlo simulations of the efficiencies of estimators. A goal of this paper is thus to further its utility. One of the drawbacks to determining the ARE of an estimator is the analytic derivation of the asymptotic variance; however, an expansion using the influence function easily yields the needed asymptotic variances.

Notation and Form of the Equation Defining the MLE

As usual, denote the response vector in an *n*-item test as $\underset{\sim}{x} = (x_1, \ldots, x_n)$; the value of the ability parameter will be denoted by θ. For a given θ, the probabilities of correct and incorrect response to the *i*th item are $P_i(\theta)$ and $Q_i(\theta)$, where $P_i(\theta) + Q_i(\theta) = 1$.

Under the assumptions of local independence, the likelihood function is

$$L(\theta) = \Pi\ [P_i(\theta)]^{x_i}\ [Q_i(\theta)]^{1-x_i}\ . \qquad [1]$$

After taking logarithms and the first derivatives with respect to θ, assuming derivatives exist, the normal equation defining the MLE is

$$\Sigma\ \frac{P_i'}{P_iQ_i}(\theta)[x_i - P_i(\theta)] = 0\ . \qquad [2]$$

For the 2-parameter logistic model,

$$[P_i(\theta)/Q_i(\theta)] = a_i(\theta - b_i) \qquad [3]$$

and

$$P_i'/(P_iQ_i) = a_i\ , \qquad [4]$$

where a_i is the discrimination parameter and b_i is the difficulty parameter of the *i*th item. The normal equation takes the form

$$\Sigma\ a_i[x_i - P_i(\theta)] = 0\ . \qquad [5]$$

Example of the Effect of an Outlier

In order to modify the MLE to cope with outliers, there must be a well-defined notion of outliers and an understanding of their effects. For example, let $n = 10$, $b_i = -0.8$ to 1.0 by steps of 0.2, and $a_i = 1$. Table 1 shows two 10-item response vectors (one without and one with an outlier at the 10th response) and their MLEs, where the items have been ordered in increasing difficulty. Apparently, the correct response to the most difficult item unwarrantedly pulls the value of the MLE by almost one-quarter of the range of the difficulty parameters; thus, the single outlier has a substantial effect on the MLE.

Table 1
MLEs for Two
10-Item Response
Vectors

$\underset{\sim}{x}$	MLE
1111100000	.10
1111100001	.58

Further inspection of the shape of the likelihood function for the second response vector in Table 1 would reveal that it is very flat in the region of the extrema; this is one way the likelihood has of warning that the value of the MLE should be treated very cautiously. However, there cannot be misgivings about the value of an estimator when immediate action is called for based on that value, especially for the computerized adaptive testing environment, which requires selection of future items based on current estimates of an examinee's ability. Therefore, estimators that can cope with outliers must be developed by reducing their effect to give estimates that are not too perturbed from estimates yielded by outlier free response vectors.

Definition of an Outlier Pair

The nature of an outlier, as demonstrated in the above example, is revealed only in reference to the value of the ability and the item characteristic function. Thus, the definition of an outlier requires the inclusion of the item characteristic function. Define $[x_i, P_i(\theta)]$ to be an outlier pair whenever $x_i = 1$ but $P_i(\theta) \approx 0$ or whenever $x_i = 0$ but $P_i(\theta) \approx 1$. Note that whenever $[x_i, P_i(\theta)]$ is an outlier, $P_i(\theta)Q_i(\theta) \approx 0$. The next definition of a new estimator will take advantage of this fact.

A New Estimator

Define an estimator $\hat{\theta}$ by the equation

$$\Sigma a_i[x_i - P_i(\hat{\theta})]\ [P_i(\hat{\theta})\ Q_i(\hat{\theta})] = 0\ . \qquad [6]$$

To understand this modification of the normal equation for the MLE, the example will be continued. Table 2 shows the two θ estimates for the same item response vectors. Thus, the PQ factor weights down the outlier by taking account of the reference value of $P_i(\theta)$ for the 10th item. This idea is continued in the following section.

Table 2
MLE and Modified
Estimate for Two
10-Item Response Vectors

$\underset{\sim}{x}$	MLE	$\hat{\theta}$
1111100000	.10	.10
1111100001	.58	.22

The h Estimators

An entire class of estimators is available for each value of h in the following equation:

$$\Sigma a_i[x_i - P_i(\theta)]\ [P_i(\theta)\ Q_i(\theta)]^{h-1} = 0 \qquad [7]$$

where a_i are real and arbitrary, but given, constants and $h \geq 1$. For $h = 1$, the equation yields the MLE. Once again, the example is continued in Table 3 to observe the effect of this modification.

Table 3
h Estimates for
Two 10-Item Response Vectors

	h				
$\underset{\sim}{x}$	1(MLE)	1.5	2	3	4
1111100000	.10	.10	.10	.10	.10
1111100001	.58	.41	.22	.20	.19

The h estimators have the following properties:

1. These estimators do not modify the MLE when the outlier is absent, since whenever $x_i - P_i(\theta)$ is small, the factor $P_i(\theta) \times Q_i(\theta)$ has little effect on the estimator.
2. This class of estimators is also available for any item response model, not just the 2-parameter logistic model. The normal equations have the form

$$\Sigma \frac{P_i'}{P_iQ_i}(\theta)[x_i - P_i(\theta)]\ [P_iQ_i(\theta)]^{h-1} = 0 . \qquad [8]$$

3. The asymptotic variance of the estimator is found by differentiating the normal equation, a step similar to taking the second derivative of the log-likelihood function, but this step is not justified by the analogy. The known asymptotic behavior of the MLE depends on the individual observations being independently and identically distributed, a condition that does not generally hold in IRT. The success of extending robustness to IRT is based on extending this condition to IRT in a natural way. This has been accomplished in Jones (1982), where the stochastic structure of IRT has been imbedded in a framework that yields independently and identically distributed observations. Although a simple concept, the precise details of the imbedding will not be explained here; however, some of the tools, such as the limit theorems for the empirical distribution function are briefly discussed as they are used in the following sections.

The Influence Function

Before defining what is meant by the term influence function, an explanation of why they are needed is in order. First, view the ability h estimator as a function of n observations, where each observation is a triplet denoted by z_i, consisting of (1) the response x_i; (2) the item characteristic function $P_i(t)$, $-\infty < t < \infty$; and (3) the constant a_i. The normal equation, Equation 7, for the h estimator is a sum whose summands are specific combinations of the components of the triplet observations. If F_n denotes the empirical distribution function of the n triplet observations (z_i)--that is, F_n gives point mass 1/n to each z_i and zero mass elsewhere--then the normal equation, Equation 7, may be written as the following Lebesque-Stieltjes integral:

$$\int g(z,\theta)\ dF_n = 0 \tag{9}$$

where

$$g(z,t) = a[x - P(t)][PQ(t)]^{h-1}\ . \tag{10}$$

By introducing the above integral, it has been demonstrated that the ability estimator θ is a function of F_n, $\theta(F_n)$, defined implicitly in Equation 9. Furthermore, the integral equation will accept any distribution function, not just F_n. Therefore, the domain of the estimator θ has also been extended to any F; thus $\theta = \theta(F)$. The reason this has been done is to facilitate the introduction of the influence function and eventually to derive its form.

The goal of these developments is to derive a useful approximation to the change of the value of an estimator attributable to the introduction of one more observation. If the observation also happens to be an outlier, then it can be determined how resistant an estimator is to the outlier. The change in the value of an estimator can be easily defined. Let F_n and F_{n+1} be the empirical distributions for n observations and the same n observations plus one more observation. The change in the value of the estimator due to the additional observation is simply

$$\theta(F_{n+1}) - \theta(F_n)\ . \tag{11}$$

The two distribution functions are close enough to allow an approximation to Equation 11 by the mean value theorem. Let $\delta(z,z_{n+1})$ be the point mass at z_{n+1}, then

$$F_{n+1}(z) = (1-s)F_n(z) + s\ \delta(z,z_{n+1})\ , \tag{12}$$

where $s = 1/(n+1)$. Since $s \to 0$ for increasing sample size, the derivative useful here has been shown in Jones (1982) to be the derivative of $\theta[(1-s)F_n + s\delta_{n+1}]$ with respect to s and evaluated at $s = 0$. Denoting this derivative by $\dot\theta(F_n;\ z_{n+1})$ gives the needed approximation to Equation 11:

$$\theta(F_{n+1}) - \theta(F_n) \simeq \dot\theta(F_n;\ z_{n+1})/(n+1)\ . \tag{13}$$

When viewed as a function of z_{n+1}, the derivative in Equation 13 is called the influence function of the estimator. Jones (1982) has shown the influence function of the estimators, defined by normal equations of the form Equation 9, to be

$$\dot\theta(F_n;z) = g(z,\theta)\Big/\frac{d}{dt}\int g(z,t)dF_n\Big|_{t=\theta}\ ; \tag{14}$$

and for the h estimator, this has the form for each observation z_i

$$\dot\theta(F_n;z_i) = \frac{[x_i-P_i(\theta)]\ w(\theta;z_i,h)}{\frac{d}{dt}\int [x-P(t)]\ w(t;Z,h)dF\Big|_{t=\theta}} \tag{15}$$

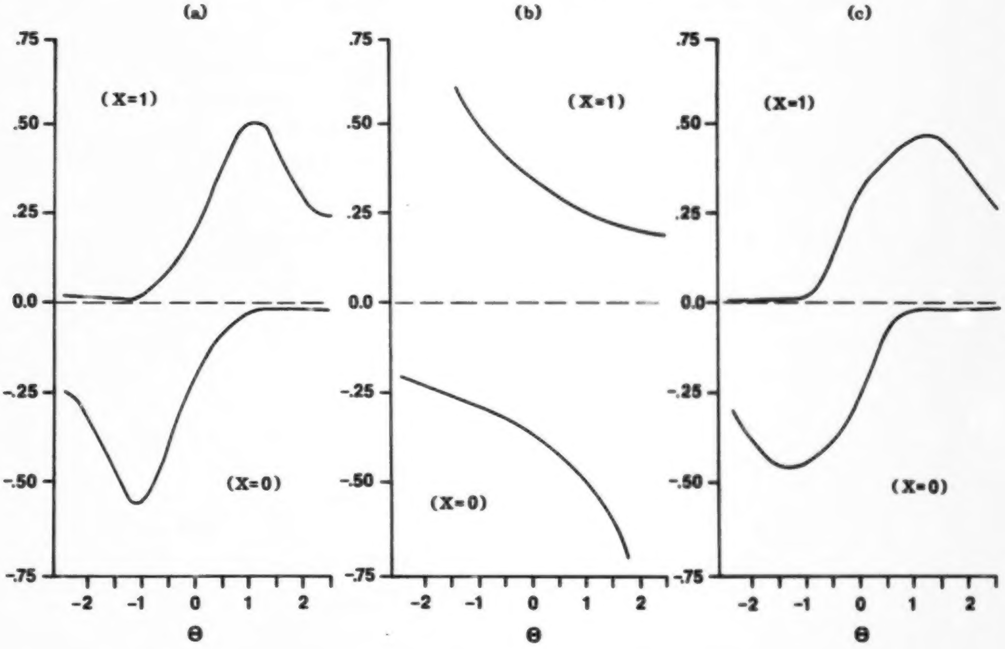
where

$$w(t;z_i,h) = a_i[P_i(t)\ Q_i(t)]^{h-1}\ . \qquad [16]$$

Graphs of Influence Functions: Influence Curves

Since the influence function is a function in $\underline{z}$, a triplet variable, it cannot simply be represented by a single graph of the mapping of the real line. However, if viewed as a function of θ when x_i is either 0 or 1, then it can be conveniently represented by two graphs of the mapping of the ability interval. Figures 1a and 1b contain the graphs of the influence functions (which will be called the influence curves) of two h estimators for h = 4 and for h = 1. When h = 1, the h estimator is just the MLE. The top curves in Figure 1 (x = 1) correspond to the influence of a correct response to the most difficult item of 10 items; the bottom curves (x = 0) are for an incorrect response to the easiest of the items.

Figure 1
Influence Curves for h Estimators with (a) h = 4,
(b) 2-Parameter MLE, and (c) Biweight Estimator with c = 2



From the figures it can be seen that for both h = 4 and h = 1, when x = 1 (a correct response) and θ is the true ability, the observations would have little effect on the estimator when the value of θ is close to 2. However, when θ is close to -2, the observation would have a very large effect for h = 1 (that is, the MLE) but little effect for an h estimator when h = 4. These analytic circumspections correspond precisely to the conclusons indicated by the examples of the previous section. How much of an effect a correct response has to the most difficult item for a person of low ability can be obtained directly from Figures 1a and 1b. The values of the influence curves are approximately 0 for the h estimator and .80 for the MLE; this latter value is about 40% of the total ability scale. Figure 1c will be discussed below.

Figure 2
Influence Curves for h Estimator with (a) h = 10
and (b) Trimmed 2-Parameter MLE

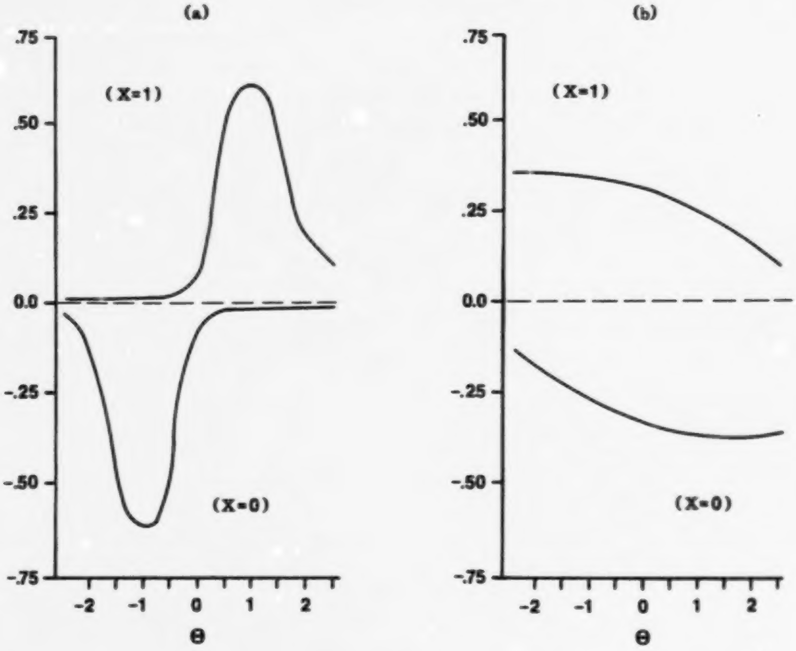


Figure 2a shows the general shape of the influence curves where h is large, e.g., h = 10. The estimator for even larger h tends to treat both correct and incorrect responses equivalently while allowing only responses to items for which the probability of a correct response is close to .5 to affect the estimator. Figure 2b is based on a new estimator discussed below.

The Properties Yielded by Influence Functions

Besides yielding an analysis of the effect of a single observation on the value of an estimator, influence functions also yield the following: (1) new robust estimators, (2) the asymptotic distributions of the estimators, and (3) comparison between competing estimators.

Derivations of New Estimators That Are Robust

For new robust estimators, new estimators can be derived simply by specifying the graph of the influence function. Figure 2b is an example providing a

MLE for correct responses when $\theta > 0$ and for incorrect responses when $\theta < 0$; however, it differs from the MLE for correct responses where $\theta < 0$ by giving less influence to the observation for each value of $\theta < 0$. The estimator behaves similarly for incorrect responses.

Asymptotic Distributions of Estimators

For the asymptotic distributions of estimators, Jones (1982) exploited the influence function approximation to the *h* estimator in order to derive its limiting distribution. If the items are randomly sampled according to a fixed distribution, then F_n will converge to a limiting distribution F (Jones, 1982). The parameters of the limiting distribution of the sequence of the estimators depend on the F, as the limiting distribution is normal with mean $\theta(F)$ and variance

$$V(F,\theta;h) = \frac{\int[x-P(\theta)]^2 \; w^2(\theta;z,h)dF}{\left[\frac{d}{dt} \int[x-P(t)] \; w(t;Z,h)dF\Big|_{t=\theta}\right]^2} \qquad [17]$$

Comparisons Between Competing Estimators

Finally, comparison between competing estimators includes (1) computing the asymptotic relative efficiency and (2) visual comparison of the influence graphs.

Asymptotic relative efficiency. The usefulness of the asymptotic relative efficiency will be demonstrated for the *h* estimators relative to the MLE. The efficiency is defined for unbiased estimators as the ratio of the asymptotic variances

$$\text{Eff} = V(F,\theta;h)/V(F,\theta;1) \; , \qquad [18]$$

and for biased estimators as the ratio of the mean squared errors (the formulas are given in Jones, 1982).

The model under which the estimators are compared correspond to the model in the example of the previous sections. The value of the efficiencies are shown in Table 4. The greatest loss of efficiency by the *h* estimators is only one item in 10 at h = 5.

Whereas the *h* estimator is always less efficient than the MLE under the assumed model, the opposite situation can occur under an alternative model. Let the alternative model be generated by perturbing the item response curves by slightly altering the item difficulty parameters; that is, for each item let the item response curve be equal to

$$P_i^*(t) = P_i(t-.1) \; . \qquad [19]$$

The resulting efficiencies are also shown in Table 4. Thus, as Table 4 shows, the efficiency of the *h* estimator can be far superior to the MLE under a model different from the assumed one.

Table 4
Values of Efficiency for Five
Levels of *h* under Assumed
and Alternative Models

Model	h 1.5	 2.0	 3.0	 4.0	 5.0
Assumed	.99	.98	.95	.92	.90
Alternative	2.24	4.83	19.15	51.46	84.62

Comparison among the biweight, Bayes, 3-parameter MLE, and the h estimators. Visual comparison of influence graphs of estimators is highly enlightening. The influence graphs of three estimators--the Bock-Mislevy biweight estimator (Mislevy & Bock, 1981), the Bayes posterior mode, and the MLE for the 3-parameter logistic model--will be illustrated.

The influence function of the biweight and MLE for the 3-parameter logistic model is the same as Equation 15 except that the form of the *w* function for the biweight is

$$w(t;z_i) = (1-t^2)^2 \quad \text{for } |t| \leq c$$

$$= 0 \quad \text{otherwise} . \qquad [20]$$

and for the 3-parameter logistic is

$$w(t;z_i) = \frac{R_i(t)}{c_i + (1-c_i)\, R_i(t)} , \qquad [21]$$

where $R_i(t)$ is a 2-parameter logistic curve and $P_i(t) = c_i + (1-c_i)\, R_i(t)$. The influence function of the Bayes posterior mode with prior $\pi(t)$ is

$$\dot{\theta}(F;z_i) = \frac{u(\theta,z_i)[x_i - P_i(\theta)]}{\frac{d}{dt} \int u(t,z)[x - P(t)]dF\big|_{t=\theta} + \psi'(\theta)} \qquad [22]$$

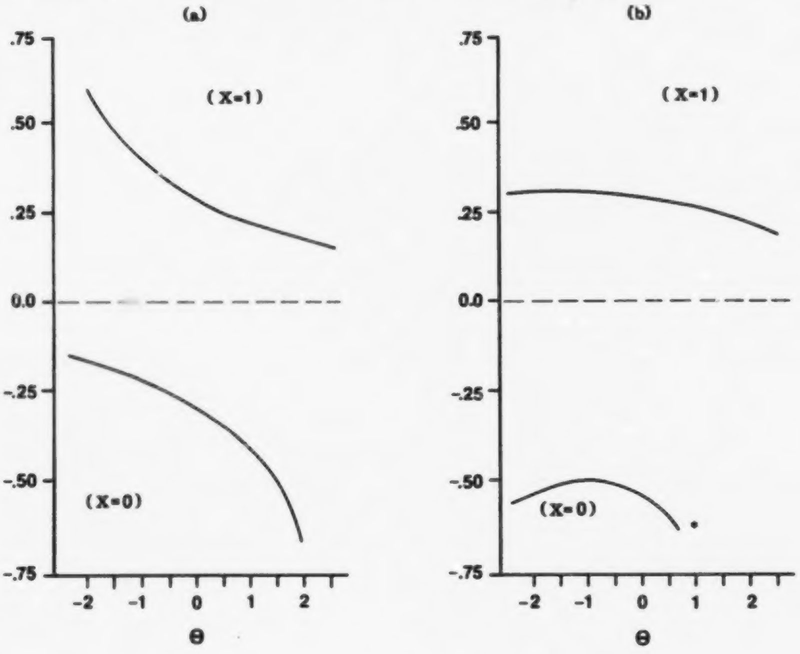
where

$$\psi(t) = \pi'(t)/\pi(t) \qquad [23]$$

and

$$u(t,z) = P'(t)/PQ(t) . \qquad [24]$$

Figure 3
Influence Curves for Posterior Mode Estimator of (a) 2-Parameter Logistic Likelihood and Normal Prior with Mean 0 and Standard Deviation 5, and (b) 3-Parameter MLE with Guessing Parameter .2



Figures 1c, 3a, and 3b illustrate the influence curves of all three estimators. It can be seen that the biweight estimator with c = 2 behaves similarly to the h estimator except that the biweight allows influence for a smaller interval of ability and, for the interval, the influence it allows is slightly larger than the h estimator. The Bayes posterior modal estimator is very similar to the 2-parameter MLE except that it uniformly downweights observations; it is surprising that there is any downweighting at all, since the standard deviation of the prior is 5 for this example. The 3-parameter MLE estimator has the most surprising influence curve of all. For correct responses, it is similar to the 2-parameter MLE, but it upweights observations with ability near the upper extreme; similarly, for incorrect responses it inflates the influence over the 2-parameter MLE, especially for large abilities. This observation is suprising because the 3-parameter MLE has been maligned as being unreliable, since the item characteristic parameters are difficult to estimate accurately. The influence curve shows, however, that the 3-parameter MLE possesses intrinsic and probably accidental robustness against outliers for correct responses, although at the expense of some undesirable sensitivity to incorrect responses. The asterisk in Figure 3b indicates that the influence curve continues off the graph, ending at -1.75 for θ = 2.

REFERENCES

Huber, P. J. *Robust statistics*. New York: Wiley, 1981.

Jones, D. H. *Redescending M-type estimators of latent ability* (ETS RB-82-30). Princeton NJ: Educational Testing Service, 1982.

Mislevy, R. J., & Bock, R. D. *Biweight estimates of latent ability*. Unpublished manuscript, 1981.

DISCUSSION

DAVID THISSEN
UNIVERSITY OF KANSAS

I am to discuss three papers which have in common only some aspects of "robustness." Presenting an integrated discussion will be difficult, but I shall make an attempt. I would first note that it is not clear that robustness is unambiguously attractive or virtuous. Is robustness monotonically attractive? I'm not sure that it is. Is all robustness in statistics virtuous? I doubt it. I will say more on that later.

I will discuss the papers in the reverse order of their presentation. First, there is "Robustness I: Robustness in Scoring Tests"--which is, however, not the title of Jones' paper on robust estimation of θ. There have been several attempts historically to solve this problem; Jones has presented the most recent and probably the best solution. Before commenting further, I would like to briefly restate the problem.

If there is a collection of calibrated test items, the virtue of item response theory (IRT) is that the difficulty of the items is supposed to be known. That may be the big difference between IRT and classical test theory. Then, if there are two people, and the first person gets the easy half of the items correct and the difficult half wrong (which is a sensible thing for a person to do), and the second person gets one less than half of the easy ones correct and most of the difficult ones wrong and then gets the most difficult one correct, there may be a problem. Using the one-parameter logistic model (which gives simpleminded statements about test scores), no matter how it is done, one claims that these two people have the same ability, which is probably silly. It is not nearly as silly, though, as the fact that both of those estimates have the same standard errors as well.

Yet, in looking at the information, it seems to me that in giving the two people the same score one is not using everything one knows. The problem is not removed by most of the rest of IRT. A Bayesian prior rescales but does not otherwise help anything, adding slopes and asymptotes scrambles things around a little bit, but there is no essential change. So the problem is that when the model does not fit, it can be harmful; and the model never fits. When people give item responses that they should not give, that is bad. There is a tradition in statistics of solving problems like this, starting with the Princeton "Robustness Study" (Andrews, Bickel, Hampel, Huber, Rogers, & Tukey, 1972) and given theoretical explication by Huber's book (1981) called Robust Statistics and in dozens of articles in the statistical literature.

Jones' estimator is a solution to the IRT problem arising from the statistical theory. It is far superior in this respect to previous ad hoc solutions to the problem such as that proposed by Wainer and Wright (1980) called AMJACK. The *h* estimators are derived from statistical theory, and can be proved to be both robust and well behaved. It is very important for us to use statistical theory periodically. This may be the best estimator we will have for a long time for scoring tests in this context.

People who score short tests with unpredictable sets of items, otherwise known as people who test with computers, probably would be well advised to try this if for no other reason than to keep computerized systems out of those traps they get into when someone either makes a lucky guess early in the adaptive test or gets something wrong that they should not. Adaptive systems can have trouble backing away from the "error." This kind of procedure used in scoring could make the system back away faster from the model's silly response.

In any event, Wainer and I are currently referees for a contest among several robust estimators to see which will turn out to be best under abusive conditions. We will tell you all about that at some future date.

I now switch topics to something nearly unrelated; but it has the word "robust" in its introduction, so it will be "Robustness II: Robustness in Parameter Estimation." There are a number of interesting results in Drasgow and Parsons' paper about multidimensionality and LOGIST, and I was more pleased about some of them than about others. I was most pleased about the fact that they could not sensibly recover the parameters for highly multidimensional tests. The slopes seemed to vary a great deal; I hope that meant one could tell one had something which was not particularly unidimensional, and that something else should be done. When the model is very wrong, the best thing the parameter estimation scheme can do is blow up. The user then won't go ahead and do anything with the parameters.

They found other results for moderately multidimensional tests. With such tests, as Drasgow and Parsons have constructed them, they have recovered the "first factor" of the domain dimension: the one line through the multidimensional space which was pretty close to the dimension of the generating space that had most items on it. There is a sense in which it could be said that this is "good," that it means this program works well, and that is is "robust" in some sense: It gives answers although the model is wrong. This is not the *error* model being wrong, as in Jones' problem, however; this is the *structural* model being wrong. It may not be desirable to have robustness against *that*.

Embretson has been complaining for years that a considerable loss could be suffered in getting only the first dimension when IRT is applied to a moderately multidimensional test. There are many tests in which the multidimensionality in the set of items is in some sense useful, even if not explicit or deliberate. That is, the set of items and its multidimensionality is really there to provide an implicit composite of several traits which are being measured, effectively weighted by the number of items on that dimension in the test; and the composite is a better predictor of something than any of the single traits. In taking something like that, getting an item response model to fit (or at least parameter estimates which look good) and then scoring it with θ on one dimension, one may be losing an important part of whatever was in the test. It would be an embarrassing situation for the raw score on the unanalyzed test to predict something better (or to be more "valid") than the result of IRT scoring; but that is possible in such a situation because of the information lost in the rejected dimensions. I am actually less pleased with the fact that numbers that look good are obtained from LOGIST for something with moderate dimensionality than I was pleased with the fact that it didn't do very well at all in high dimensionality. This kind of "robustness" may be bad.

Things could even be worse. If, unknowingly, one had a multidimensional item set, the situation could be such that dimensional identity of an item was correlated with its difficulty. For instance, in a "verbal" test there could be vocabulary items which are a good deal more difficult than reading comprehension items, so that the test is somewhat multidimensional and the multidimensionality correlated with the locations of the items. We know from Drasgow and Parsons' results that we can't tell with the most widely used IRT program whether the test is multidimensional. An adaptive test constructed from such an item pool could "adapt" to the high ability people by measuring them on the dimension with the more difficult items; and to the lower ability people, by measuring them on the other dimension with the easier items. This would miraculously (and invisibly) turn a test that was supposed to measure one thing into a test measuring different things for different people. I'm not sure how well or badly that might work; but I doubt that it is a virtue.

So one thing we would really like to know is how to tell what the dimensionality of a test is. It is clear that estimating the parameters of an item response model, at least in the ways we do it now, will probably not tell us that. We should therefore look with interest on the first paper of this trio by Stout. He describes a statistic which is intended to do just that: Point the statistic at the test, and it is supposed to tell you the dimensionality. As I've tried to indicate above, that could be very important.

In some sense, in principle, all that is obtained from a set of test data is a contingency table in which each examinee is in a cell of a 2^k table ($\underline{k}$ = number of items). Again, in principle, when we say these models fit, we mean there is supposed to be independence in that table, conditional on some underlying unidimensional latent variable. This can be investigated by chi-square analysis of the table. The trouble with that set of principles is that for 20 items it is going to be a 2 table, which has somewhat more than one million cells. For 40 items there are a million million cells. Contemporary statistical theory does not include distributional theory for a million million cells.

Stout's statistic may solve this problem. If so, it will fill an important gap. Of course, at this point, there is more question than answer: since he has given us a formula for the statistic, the first question is, "will it work?" There is more question here than is obvious because it is explicitly an asymptotically valid statistic, but nobody gives tests of asymptotic length. People give 20-item tests. So, a second question is "will it work for short tests?" Stout reports that they are working on answers to those questions; and we will look forward to those answers.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. Robust estimates of location. Princeton, NJ: Princeton University Press, 1972.

Huber, P. J. Robust statistics. New York: Wiley, 1981.

Wainer, H. & Wright, B. Robust estimation of ability in the Rasch model. Psychometrika, 1980, 45, 373-391.

ROBUSTNESS OF ADAPTIVE TESTING TO MULTIDIMENSIONALITY

DAVID J. WEISS AND DEBRA SUHADOLNIK
UNIVERSITY OF MINNESOTA

Before computerized adaptive testing (CAT) can be applied in various operational settings, its characteristics must be evaluated under a variety of conditions. Studies of the reliability and validity of CAT (e. g., Johnson & Weiss, 1980; Kiely, Zara, & Weiss, 1983; McBride & Martin, 1983; Moreno, Wetzel, McBride & Weiss, 1984; Sympson, Weiss, & Ree, 1982) provide important information comparing CAT to conventional tests in applied situations. Live-testing studies such as these, however, are expensive and time-consuming, provide results that are dependent on the characteristics of the sample of subjects and the specific criterion variables used, and do not permit an answer to the important questions about how well CAT measures true ability levels and whether ability is better estimated at different ability levels. Live-testing studies also incorporate a number of uncontrolled sources of error (e. g., item parameter estimation error, various errors of measurement due to idiosyncratic characteristics of examinee responses to test items) which further complicate the process of reaching generalizable conclusions.

Monte carlo simulation provides a means of systematically examining the performance of CAT under a variety of conditions and of identifying the effects of various kinds of errors on the performance of CAT strategies. Early studies were concerned with the comparison of CAT item selection strategies with conventional tests (e.g., Betz & Weiss, 1973, 1974, 1975; Larkin & Weiss, 1974; Vale & Weiss, 1975a, 1975b) and with each other (e.g., Larkin & Weiss, 1975). These studies provided global evaluations of CAT strategies that were useful in eliminating some strategies from further consideration. Later studies then concentrated on the more promising strategies, generally those that are based on item response theory (IRT), examining the performance of these testing strategies conditional on true ability levels (e.g., McBride, 1977; Vale, 1975; Weiss & McBride, 1984).

One factor that can affect the performance of CAT is the nature of the item pool from which it draws items. McBride (1977; Weiss & McBride, 1984) studied the performance of a Bayesian CAT in perfect and ideal item pools and in realistic item pools in which the IRT diffficulty and discrimination parameters were correlated. Others (e.g., Urry, 1974) also examined CAT performance in a variety of item pool configurations.

In addition to the distributions of item difficulties and discriminations in a given item pool, the degree of error in the IRT item parameter estimates in a real item pool can affect the performance of CAT, particularly since items are selected on the basis of their IRT parameter estimates. Crichton (1981) investigated the effects of errors in item parameter estimates on the performance of maximum information and Bayesian CAT strategies in the context of the 3-parameter logistic model. Mattson (1983) extended Crichton's study to the 1- and 2-parameter logistic models, both Bayesian and maximum likelihood scoring, and to the more realistic situation in which the IRT difficulty and discrimination parameters had varying degrees of correlation. These later studies provide valuable information about the performance of CAT under the realistic situation in which adaptive testing is to be done using item pools with parameters estimated with varying degrees of error.

A second factor that is likely to have an effect on the performance of IRT-based CAT is multidimensionality. Operational IRT models used for CAT assume that unidimensionality exists at two stages: (1) in the process of estimating item parameters, and (2) in the process by which an individual generates a response to a test item with given item parameters. Presumably, any deviations from unidimensionality that exist at either of those stages in CAT could result in non-optimal performance of IRT-based CAT strategies.

While many tests of ability and achievement approximate unidimensionality, none have shown the strict unidimensionality required by operational IRT models. This motivated Drasgow and Parsons (1983) to examine the effects of deviations from unidimensionality during the item parameter estimation process on IRT item parameter estimates.

Purpose

The present monte carlo simulation study was designed to examine the effects of multidimensionality during CAT test administration. It was assumed that multidimensionality existed in the individuals to whom test items were being administered--i.e., that the correct or incorrect responses given by an individual were generated from a specified multidimensional structure, rather than the unidimensional IRT model normally assumed to have generated the observable dichotomous test item responses. The dichotomous response was then treated for CAT item selection and ability estimation purposes as if it had been generated by the unidimensional model. To the extent that the observed item response was affected by dimensions other than the first (which corresponded to the single dimension assumed to underlie the item selection and ability estimation process) errors should be introduced into the adaptive testing process. These errors should affect the ability estimates and the efficiency of CAT. The study focused on the nature and degree of these errors under a variety of multidimensional structures, to determine how robust CAT is to the effects of multidimensionality in examinees' responses to test items.

METHOD

Initial Factor Analyses

Item response vectors for forms 8A and 8B of the Armed Services Vocational

Aptitude Battery (ASVAB) were obtained for a sample of military recruits. For those subtests of the ASVAB (Mathematics Knowledge, General Science and Mechanical Comprehension) in which forms 8A and 8b were identical except for the order of the items, the response vectors for form 8B were rearranged to match the order of the items in form 8A. This resulted in datasets with sample sizes of 5,127 for these three subtests, sample sizes of 2,621 for form 8A of the other seven subtests, and sample sizes of 2,506 for form 8B of the other seven subtests.

Tetrachoric inter-item correlations were computed for eight of the ten subtests; the Numerical Operations and Coding Speed subtests were not included in further analyses due to the speeded nature of these subtests. The tetrachoric correlations for the other eight subtests were then factor-analyzed using a principal axes factor extraction method and a Varimax rotation. Of the resulting factor structures, the factor structure of the General Science subtest exhibited the greatest degree of multidimensionality. Table 1 lists the factor loadings on the first four factors for the items in this subtest. This factor structure was used as the model for generating subsequent factor structures with varying degrees of multidimensionality.

Generation of Factor Structures

The first step in creating factor structures with varying degrees of multidimensionality was to round the 25 factor loadings on the first factor of the ASVAB General Science (GS) subtest to the nearest multiple of .05. This set of 25 rounded factor loadings was then repeated six times to create a set of factor loadings for 150 items on one factor with the same configuration of loadings as the first factor for the ASVAB GS subtest. This factor, the original strength ASVAB factor (OSAF), was used as the basis for one of three sets of factor structures.

Sixteen factor structures of varying dimensionality were constructed using OSAF as the first factor. Factors other than the first factor were constructed to be proportional in strength to the first factor. These sixteen factor structures are described in Table 2. Factor structures varied from a 2-factor structure with the second factor 1/8 as strong as the first factor (Dataset 2) to a 3-factor structure with Factors 2 and 3 equal in strength to Factor 1 (Dataset 16). An additional dataset (17) consisted of the actual 4-factor ASVAB GS factor solution.

The 150 factor loadings on OSAF were then increased to yield a first factor that was approximately 1.5 times as strong as OSAF. This new first factor (1.5 OSAF) was used as the first factor in a set of sixteen different factor structures which are also described in Table 2 (Datasets 18-33). Factors other than the first factor in Datasets 18-32 were again constructed to be proportional to this strengthened first factor in all of the factor structures except the 4-factor structure (Dataset 33), where the second, third and fourth factors were the actual second, third, and fourth factors from the original factor analysis of the ASVAB GS subtest (see Table 1).

The 150 factor loadings on OSAF were then increased a second time to result

Table 1
Factor Loadings for the First Four Factors of the ASVAB
General Science Subtest

Item Number	Factor 1	Factor 2	Factor 3	Factor 4
1	.540	-.215	-.250	.027
2	.624	-.205	-.018	-.303
3	.642	-.201	-.095	.026
4	.486	-.098	-.118	-.115
5	.668	-.233	.162	.069
6	.703	-.160	.066	.073
7	.572	.052	.103	-.019
8	.493	-.070	.072	-.067
9	.546	-.174	-.239	.119
10	.547	-.212	-.015	.016
11	.595	.060	.009	-.025
12	.398	.099	.058	-.006
13	.580	.096	-.120	-.233
14	.580	-.172	-.069	.124
15	.438	-.029	.043	.337
16	.543	.012	.100	.172
17	.462	.120	-.030	-.009
18	.639	.227	.054	-.072
19	.371	.208	.045	-.011
20	.473	.048	.132	-.096
21	.460	.273	.085	.006
22	.283	.224	.115	-.032
23	.480	.035	.147	-.062
24	.387	.650	-.310	.067
25	.396	.310	.101	.089
Factor Contribution	7.541	1.671	1.030	1.023

in a first factor that was approximately twice as strong as OSAF. This strengthened first factor (2.0 OSAF) was used as the first factor in a third set of twelve factor structures (Datasets 34-45), which are also described in Table 2. In Datasets 34-44, factors other than the first factor were constructed to be proportional in strength to this increased strength first factor; these additional factors were also constructed to avoid communalities greater than 1.0 for any item. For the 4-factor structure of Dataset 45, the factors other than the first factor were taken directly from the original factor analysis of the ASVAB GS subtest (see Table 1).

Generation of Response Vectors

To evaluate the effect of violation of the assumption of unidimensionality in adaptive testing, sets of dichotomous (0,1) item responses were generated using the factor structures with varying degrees of multidimensionality.

Table 2
Dataset Numbers for Datasets Based on
First Factors of 1.0, 1.5, and 2.0 OSAF,
and Factor Strengths of Factors 2 through 4,
for Each of the Datasets

Dataset Number			Factor Strength as a Proportion of Factor 1		
1.0 OSAF	1.5 OSAF	2.0 OSAF	Factor 2	Factor 3	Factor 4
1	18	34	-	-	-
2	19	35	1/8	-	-
3	20	36	1/4	-	-
4	21	37	1/3	-	-
5	22	38	1/2	-	-
6	23	39	2/3	-	-
7	24	40	3/4	-	-
8	25	--	1.0	-	-
9	26	41	1/8	1/8	-
10	27	42	1/4	1/4	-
11	28	43	1/3	1/3	-
12	29	44	1/2	1/4	-
13	30	--	1/2	1/2	-
14	31	--	2/3	1/3	-
15	32	--	2/3	2/3	-
16	--	--	1.0	1.0	-
17	33	45	GS-2*	GS-3*	GS-4*

*Factor derived from factor analysis of ASVAB GS test.

The first step was to assign θ levels for each factor to a number of hypothetical examinees (simulees). This was done for each factor except the first factor by using a random number generator to create uniform distributions of 1,700 θ values between -3.2 and +3.2 for each factor independently of all other factors. θ levels for the first factor were assigned so that 100 simulees were assigned to each of 17 θ levels ranging from -3.2 to +3.2 in increments of .4. θ levels for the first factor were assigned in this manner in order to have a sufficient number of replications at each θ level so that indices conditional on θ could be computed.

Next, matrices of item response theory (IRT) item parameters were calculated and generated. Item discrimination parameters (*a*s) were computed using the following formula:

$$a_{gj} = F_{gj}/[1 - (F_{gj}^2)]^{\frac{1}{2}} \qquad [1]$$

where a_{gj} = item discrimination parameter for item *g* and factor *j*, and
F_{gj} = factor loading for item *g* on factor *j*.

These matrices of *a* parameters were calculated for each of the 45 factor structures.

Matrices of item difficulty parameters (*b*s) were generated for each of the 45 factor structures using a random number generator which generated a uniform distribution of 150 values between -3.2 and +3.2 independently for each factor in a given factor structure. Item pseudo-guessing parameters (*c*s) were also generated for each factor in the 45 factor structures; they were generated to yield a normal distribution of 150 values with a mean of .20 and a standard deviation of .02 for each factor.

After the item parameter matrices for each factor structure were determined, the probability of a correct response to each item for each factor was computed for each of the 1,700 simulees using the three-parameter logistic model,

$$P_{igj}(\theta_j) = c_{gj} + \frac{(1 - c_{gj})}{1 + \exp[-1.7a_{gj}(\theta_j - b_{gj})]} \qquad [2]$$

where $P_{igj}(\theta_j)$ = probability of a correct response to item *g* on factor *j* for a simulee with trait level θ_j,

c_{gj} = IRT pseudo-guessing parameter for item *g* on factor *j*,

a_{gj} = IRT discrimination parameter for item *g* on factor *j*, and

b_{gj} = IRT difficulty parameter for item *g* on factor *j*.

The probabilities for each item on each factor were then combined using Equation 3 to calculate the overall probability of a correct response for each individual on each item:

$$r_{ig} = \frac{\sum_{j=1}^{K} F_{gj}^2 P_{igj}}{\sum_{j=1}^{K} F_{gj}^2} \qquad [3]$$

where r_{ig} = overall probability of a correct response for simulee *i* on item *g*,

F_{gj} = factor loading for item *g* on factor *j*, and

P_{igj} = probability of a correct response for simulee *i* on factor *j* for item *g*.

Dichotomous item scores (u_{ig}) were then generated using r_{ig} and a random number generator. For each simulee and item, a random number between 0 and 1 was generated. If r_{ig} was greater than this random number, an item score u_{ig} = 1 was assigned for the response of simulee *i* to item *g*. If r_{ig} was less than the random number, an item score u_{ig} = 0 was assigned to the item for the simulee. In this manner, each of the 1,700 simulees received an item score of 0 or 1 on each of the 150 items for each factor structure.

Adaptive Testing Strategy

The sets of dichotomous item responses u_{ig} generated from the factor structures with varying degrees of multidimensionality were used with a maximum information adaptive testing strategy to obtain θ estimates. Since the adaptive testing strategy used assumes a unidimensional set of item responses, the obtained θ estimates can be used to determine the effect of violation of the assumption of unidimensionality. For each factor structure:

1. $\hat{\theta}$ was set to 0.0 for each simulee.

2. Information at $\hat{\theta}$ was computed for each of the 150 items using first factor *a*, *b*, and *c* parameters in the following equation:

$$I_g(\hat{\theta}) = [P'_g(\hat{\theta})]^2/P_g(\hat{\theta})Q_g(\hat{\theta}) \qquad [4]$$

where $I_g(\hat{\theta})$ = information at $\hat{\theta}$ for item *g*,
$P_g(\hat{\theta})$ = probability of a correct response to item *g* at $\hat{\theta}$,
$P'_g(\hat{\theta})$ = first derivative of $P_g(\hat{\theta})$, and
$Q_g(\hat{\theta}) = 1 - P_g(\hat{\theta})$.

3. The item with the highest level of information at $\hat{\theta}$ was selected as the next item to be administered.

4. The item responses to the item chosen to be administered were read from the generated item response matrix for each simulee.

5. A new $\hat{\theta}$ was calculated for each simulee using maximum likelihood scoring:

$$L(\theta_{i1}|u_i) = \prod_{g=1}^{K} P_{ig}(\hat{\theta}_i)^{u_{ig}}\, Q_{ig}(\hat{\theta}_i)^{1-u_{ig}} \qquad [5]$$

where $L(\theta_{i1}|u_i)$ = likelihood of the simulee's observed response pattern (u_i) at θ_{i1},
$P_{ig}(\hat{\theta}_i)$ = probability of a correct response to item *g* for simulee *i* with trait level estimate $\hat{\theta}_i$,
u_{ig} = 1 for a correct response to item *g*,
= 0 for an incorrect response to item *g*,
$Q_{ig}(\hat{\theta}_i) = 1 - P_{ig}(\hat{\theta}_i)$, and
K = the number of items administered.

The value of θ which had the greatest likelihood for the observed item responses was selected as the new θ estimate for a simulee ($\hat{\theta}$ was restricted to the range +4 to -4).

6. Steps 2 through 5 were repeated using the new $\hat{\theta}$s for each simulee until 30 items were administered;

7. The $\hat{\theta}$s were saved at 5, 10, 15, 20, 15 and 30 items.

Evaluative Indices

Conditional indices. Since no one optimal evaluative index was available, four different evaluative indices were used to determine the effect of violations of the assumption of unidimensionality in adaptive testing. Each of the following four indices were computed at each of the 17 θ levels on the first factor and for all six test lengths.

1. Bias:

$$\text{Bias}(\theta_{p1}) = \frac{\sum_{i=1}^{N(\theta_{p1})} (\hat{\theta}_i - \theta_{i1})}{N(\theta_{p1})} \qquad [6]$$

$\hat{\theta}_i$ = estimated θ level for simulee i,
θ_{p1} = true θ level for simulee i on factor 1, and
$N(\theta_{p1})$ = number of simulees at level p (usually 100, but occasionally smaller due to maximum likelihood convergence failures).

This index takes into account both the size and direction of the difference between true and estimated θ.

2. Inaccuracy:

$$\text{Inaccuracy}(\theta_{p1}) = \frac{\sum_{i=1}^{N(\theta_{p1})} \left|\theta_i - \theta_{i1}\right|}{N(\theta_{p1})} \qquad [7]$$

Inaccuracy considers only the size, and not the direction, of the difference between estimated and actual θ levels for each simulee at a given θ level and test length.

3. Root Mean Square Error (RMSE). RMSE was calculated as

$$\text{RMSE}(\theta_{p1}) = \left(\frac{\sum_{i=1}^{N(\theta_{p1})} (\hat{\theta}_i - \theta_{i1})^2}{N(\theta_{p1})} \right)^{\frac{1}{2}} \qquad [8]$$

This index gives more weight to larger differences between estimated and true θ levels.

4. Efficiency. Efficiency was defined by

$$I(\theta_1) = \frac{\sum_{g=1}^{K} I_{g*}(\theta_1)}{\sum_{g=1}^{K} I_g(\theta_1)} \qquad [9]$$

where g* indexes items actually administered and g indexes the items with the maximum levels of information at θ_1.

Thus, efficiency is the ratio of the information in the k items actually administered to the k most informative items at θ_1. It will equal 1.0 when the adaptive testing strategy administers the k items with maximum information at θ_1. Deviations from 1.0 result from the fact that, at any stage of the adaptive test, $\hat{\theta}$ is not usually exactly equal to θ_1.

Comparison of conditional multidimensional and unidimensional results. To summarize the effects of multidimensionality on each of the evaluative indices, distance measures were computed across the 17 θ_1 levels between the values of each of the conditional evaluative indices for the unidimensional (UD) datasets and the multidimensional (MD) datasets for all six test lengths. Cronbach and Gleser's (1953) formulas were used for computing a distance measure, D^2, between two profiles and for decomposing D^2 into components due to mean differences, scatter differences, and shape differences. Profiles were plots of the values of an evaluative index for a given dataset and test length across all 17 θ_1 levels. The formulas used were:

$$D^2_{UD,MD} = \sum_{p=1}^{17} (X_{pUD} - X_{pMD})^2 \qquad [10]$$

where $D^2_{UD,MD}$ = overall squared distance between profile UD and profile MD,
X_{pUD} = value of the evaluative index for dataset UD and θ level p, and
X_{pMD} = value of the evaluative index for dataset MD and θ level p.

$$D^{2'}_{UD,MD} = D^2_{UD,MD} - 17(\Delta^2 EL_{UD,MD}) \qquad [11]$$

where $D^{2'}_{UD,MD}$ = squared distance between profiles UD and MD after differences in mean level between the two profiles are eliminated.
$\Delta^2 EL_{UD,MD}$ = squared difference in mean level between profiles UD and MD, and

$$D^{2''}_{UD,MD} = \frac{D^{2'}_{UD,MD} - \Delta^2 S_{UD,MD}}{S_{UD} S_{MD}} \qquad [12]$$

where $D^{2''}_{UD,MD}$ = squared distance between profiles UD and MD after differences due to mean level and scatter between the two profiles are eliminated

$$S_{UD} = \left(\sum_{p=1}^{17} (X_{pUD} - \bar{X}_{UD})^2 \right)^{\frac{1}{2}} \qquad [13]$$

where S_{UD} = scatter for profile UD,
$\bar{X}_{UD}$ = mean of the 17 values of the evaluative index for profile UD,
$\bar{X}_{MD}$ and S_{MD} are defined similarly, and
$\Delta^2 S_{UD,MD}$ = squared difference between scatters for profiles MD and UD.

The presence of scatters less than 1.00 for many of the datasets resulted in values of $D^{2''}$ that were larger than the values of $D^{2'}$ for the same profiles. This made interpretation of the distance measures difficult, so the values of each of the four evaluative indices at each of the 17 θ levels were multiplied by 10. This fact should be taken into account in interpreting the magnitude of the differences between profiles and the distance measures.

To aid in interpreting the differences in profiles due to level, scatter and shape, the proportion of the squared distance (D^2) due to each of these components was computed using the following formulas:

$$\text{Level Effect}_{UD,MD} = \frac{D^2_{UD,MD} - D^{2'}_{UD,MD}}{D^2_{UD,MD}} , \qquad [14]$$

the proportion of D^2 due to differences in level between profiles UD and MD,

$$\text{Scatter Effect}_{UD,MD} = \frac{D^{2'}_{UD,MD} - D^{2''}_{UD,MD}}{D^2_{UD,MD}} , \qquad [15]$$

the proportion of D^2 due to differences in scatter between the two profiles, and

$$\text{Shape Effect}_{UD,MD} = \frac{D^{2''}_{UD,MD}}{D^2_{UD,MD}} , \qquad [16]$$

the proportion of D^2 due to differences in shape between profiles MD and UD.

Unconditional indices. In addition to examining the bias, inaccuracy, and RMSE conditional on θ level, mean values of these indices were computed across the 17 θ levels for each dataset and test length. Also computed for each condition was the fidelity correlation between $\hat{\theta}$ and θ_1. These correlations were computed for a normally distributed sample of 630 simulees selected from the 1,700 rectangularly distributed simulees in each dataset.

RESULTS

Unconditional Indices

Fidelity

Table 3 shows fidelity correlations for each of the datasets based on OSAF, 1.5 OSAF, and 2.0 OSAF, as a function of test length. For the single-factor Dataset 1, fidelity increased with increasing test length from .646, when 5 items were administered, to .928 at 30 items. For the 2-factor datasets (2-8) fidelity generally decreased with increasing strength of the second factor with two exceptions: (1) Dataset 5, which had a second factor 1/2 as strong as the first factor, had consistently higher fidelity than Dataset 4, in which the second factor was only 1/3 as strong as the first; and (2) Dataset 6, in which the second factor was 2/3 the strength of the first, had consistently lower fidelity than Dataset 7, in which the second factor was slightly stronger (3/4 of the first). In both these cases, differences between the fidelities decreased with increases in test length. For all datasets, fidelity increased with increasing test length.

For these 2-factor datasets, multidimensionality had fairly substantial effects on fidelity. For example, at the 15-item test length fidelity was .872 for the single-factor Dataset 1, but dropped to .548 when there were two equal factors (Dataset 8). When the second factor was only 1/4 the strength of the first factor (Dataset 3), fidelity for a 15-item test decreased from .872 to .784. To overcome the effect of this degree of multidimensionality, the 15-item test of Dataset 3 would need to be doubled in length, resulting in a fidelity of .880. For degrees of multidimensionality beyond those represented by Dataset 3, tests would need to be well beyond 30 items in length to equal the fidelity of the 15-item test in UD Dataset 1.

A similar pattern of results was observed for the 3-factor structures (Datasets 9-16), but the effects of multidimensionality on fidelity were even stronger. In these datasets there was, again, a general decrease in fidelity with increasing strength of the second and third factors. Fidelity also increased with test length for all datasets. In general, however, fidelities were lower for the 3-factor datasets than for those with two factors, even when the total variance accounted for by factors beyond the first was equal. For example, at the 15-item length, fidelity for Dataset 13 (with factors 2 and 3 each 1/2 of the first factor in strength) was .443; when the same amount of variance was concentrated in only the second factor (Dataset 8), fidelity was .548. Only Dataset 9, with second and third factors each 1/8 of the first factor, attained a sufficiently high fidelity at 30 items (.869) to approximate that of UD Dataset 1 at 15 items (.872).

Results for the 1.5 and 2.0 OSAF datasets were similar to those for 1.0 OSAF, with a general increase in fidelity with increasing strength of the first factor. For example, for a 15-item test based on a 2-factor structure with the second factor 3/4 the strength of the first factor, fidelity was .628 for 1.0 OSAF (Dataset 7), .685 for 1.5 OSAF (Dataset 24), and .789 for 2.0 OSAF (Dataset 40). For the 3-factor datasets with the second and third factors each 1/3 of

Table 3

Fidelity as a Function of Test Length for
Unidimensional (UD) and Multidimensional Datasets
Based on First Factors 1.0, 1.5, and 2.0 Times as
Strong as the ASVAB General Science Factor

Dataset	No. of Factors	Test Length (Number of Items)					
		5	10	15	20	25	30
1.0 OSAF							
1 (UD)	1	.646	.799	.872	.903	.914	.928
2	2	.592	.762	.823	.866	.896	.909
3	2	.519	.692	.784	.833	.863	.880
4	2	.461	.592	.672	.718	.765	.790
5	2	.534	.648	.711	.780	.813	.826
6	2	.404	.543	.616	.658	.677	.705
7	2	.431	.572	.628	.662	.694	.715
8	2	.429	.510	.548	.580	.616	.631
9	3	.522	.665	.760	.821	.847	.869
10	3	.423	.567	.655	.706	.737	.763
11	3	.375	.477	.567	.633	.678	.710
12	3	.340	.467	.559	.614	.652	.679
13	3	.320	.386	.443	.499	.548	.584
14	3	.350	.467	.529	.574	.618	.645
15	3	.313	.383	.418	.434	.473	.490
16	3	.267	.339	.371	.400	.415	.438
17	4	.577	.723	.802	.847	.871	.893
1.5 OSAF							
18 (UD)	1	.691	.842	.916	.937	.949	.955
19	2	.660	.822	.881	.912	.931	.945
20	2	.587	.753	.848	.892	.914	.924
21	2	.560	.740	.828	.877	.904	.912
22	2	.569	.737	.808	.842	.867	.878
23	2	.462	.616	.724	.772	.802	.812
24	2	.478	.623	.685	.713	.748	.763
25	2	.387	.510	.607	.651	.675	.697
26	3	.590	.740	.816	.863	.892	.911
27	3	.446	.596	.702	.752	.782	.801
28	3	.439	.569	.654	.710	.755	.776
29	3	.442	.578	.650	.702	.742	.759
30	3	.447	.554	.637	.695	.731	.745
31	3	.455	.589	.690	.732	.756	.771
32	3	.415	.525	.610	.653	.681	.700
33	4	.581	.765	.858	.892	.918	.932
2.0 OSAF							
34 (UD)	1	.733	.867	.930	.953	.961	.965
35	2	.585	.775	.888	.932	.955	.964
36	2	.599	.749	.850	.911	.927	.937
37	2	.524	.694	.817	.860	.902	.924
38	2	.604	.710	.807	.853	.866	.888
39	2	.547	.655	.756	.816	.843	.849
40	2	.542	.689	.789	.813	.836	.844
41	3	.519	.690	.804	.875	.923	.929
42	3	.542	.647	.744	.813	.841	.868
43	3	.499	.631	.758	.831	.855	.874
44	3	.534	.674	.777	.819	.840	.857
45	4	.379	.482	.546	.618	.664	.700

the first factor, fidelity for a 15-item test in the 1.0 OSAF data was .567 (Dataset 11), rising to .654 when factor 1 was 1.5 OSAF (Dataset 28) and to .758 with 2.0 OSAF (Dataset 43). As in the 1.0 OSAF data, a single factor beyond the first had less effect on fidelity than did two factors equaling the strength of the single factor, though the effect diminished substantially with the stronger first factor. For example, in the 1.5 OSAF structures for a 15-item test with a second factor 2/3 of the first factor (Dataset 23), fidelity was .724 versus .654 when there were two factors beyond the first, each comprising 1/3 of the first factor (Dataset 28); comparable factor structures with 2.0 OSAF resulted in fidelities of .756 (Dataset 39) and .758 (Dataset 43).

Datasets 17, 33, and 45 provide results based on factors derived from the ASVAB 4-factor structure, in which factors 2, 3, and 4 accounted for 22.2%, 13.6%, and 13.5%, respectively, of OSAF. Table 3 shows that there were relatively small effects on fidelity for the 1.0 and 1.5 OSAF datasets, particularly for tests of 20 or more items. For example, in Dataset 17 fidelity for a 25-item test was .871 versus .914 for UD Dataset 1. Comparable results for the 1.5 OSAF data were .918 (Dataset 33) and .949 (Dataset 18). In the 2.0 OSAF data, however, the 4-factor ASVAB structure (Dataset 45) resulted in the lowest observed fidelities for those datasets; fidelity dropped from .953 (UD Dataset 34) to .618 for ASVAB at 20 items, and from .965 to .700 at 30 items.

Bias, Inaccuracy, RMSE

Table 4 provides data on mean bias, inaccuracy, and RMSE for the datasets based on 1.0 OSAF. For UD Dataset 1, bias decreased from .282 at 5 items to .010 at 30 items. Each of the 2-factor datasets (2-8) showed lower levels of positive bias and higher levels of negative bias than did Dataset 1, with bias becoming increasingly negative as the strength of the second factor increased. Thus, in 2-factor data structures $\hat{\theta}$ underestimated θ, on the average, as both test length and strength of multidimensionality increased. A similar trend was observed for most of the 3-factor datasets (9-16), with a few exceptions. In these datasets bias tended to become less positive and increasingly negative for all test lengths for Datasets 9-12, in which the sum of the variance accounted for by the second and third factors was less than that of the first factor. In Dataset 13, which had second and third factors each 1/2 of the first factor, bias was again positive for tests of 15 items or less, but this effect was reversed for Dataset 14 (factor 2 = 2/3 of factor 1, and factor 3 = 1/3 of factor 1). However, for tests of 5 or 10 items, bias then again became positive for Datasets 15 and 16, which had very strong second and third factors. There was also a slight trend toward positive mean bias in Dataset 16. As Table 4 also shows, there was a slight effect on bias when data were generated from the 4-factor ASVAB structure (Dataset 17). For these data the ASVAB structure resulted in a slight mean underestimation of θ at test lengths of 20 to 30 items with a mean bias of .006 at 15 items compared with .038 for Dataset 1.

Both inaccuracy and RMSE tended to increase with increasing strength of factors beyond the first, and to decrease with increasing test length; this held true for both the 2- and 3-factor datasets. An exception occurred for Dataset 14 (factor 2 = 2/3 of factor 1, and factor 3 = 1/3 of factor 1) for both inaccuracy and RMSE. For this dataset inaccuracy and RMSE values were lower than

Table 4

Mean Bias, Inaccuracy, and RMSE as a Function of Test Length for Unidimensional (UD) and Multidimensional Datasets, Based on ASVAB General Science Factor

Dataset	No. of Factors	Test Length (Number of Items) 5	10	15	20	25	30
Bias							
1 (UD)	1	.282	.107	.038	.024	.015	.010
2	2	.247	.100	.031	-.015	-.028	-.026
3	2	.163	.056	-.004	-.026	-.042	-.051
4	2	.189	.060	-.022	-.052	-.072	-.084
5	2	.164	.065	-.017	-.053	-.075	-.085
6	2	.170	.038	-.023	-.070	-.099	-.107
7	2	.023	-.071	-.125	-.136	-.147	-.153
8	2	.057	-.026	-.103	-.135	-.171	-.190
9	3	.306	.133	.033	.012	-.020	-.028
10	3	.113	-.007	-.080	-.090	-.101	-.115
11	3	.173	.007	-.039	-.086	-.104	-.115
12	3	.128	.022	-.046	-.068	-.096	-.111
13	3	.397	.231	.061	-.040	-.089	-.131
14	3	-.051	-.097	-.127	-.174	-.193	-.201
15	3	.139	.059	-.060	-.122	-.171	-.210
16	3	.379	.240	.101	.006	-.068	-.142
17	4	.214	.075	.006	-.028	-.034	-.032
Inaccuracy							
1 (UD)	1	.906	.587	.451	.388	.357	.332
2	2	.982	.657	.517	.446	.402	.370
3	2	1.055	.713	.570	.493	.447	.413
4	2	1.251	.941	.770	.676	.617	.573
5	2	1.183	.870	.715	.604	.549	.521
6	2	1.247	.948	.791	.713	.664	.638
7	2	1.308	1.012	.861	.789	.733	.698
8	2	1.387	1.112	.997	.915	.863	.841
9	3	1.146	.781	.603	.512	.455	.418
10	3	1.373	1.027	.848	.731	.661	.612
11	3	1.424	1.100	.915	.801	.725	.675
12	3	1.455	1.135	.948	.837	.765	.720
13	3	1.622	1.292	1.113	1.016	.941	.888
14	3	1.549	1.244	1.062	.954	.875	.829
15	3	1.643	1.419	1.298	1.200	1.131	1.092
16	3	1.733	1.492	1.371	1.280	1.222	1.179
17	4	1.055	.734	.581	.489	.435	.399
RMSE							
1 (UD)	1	1.211	.785	.603	.514	.461	.425
2	2	1.328	.904	.694	.591	.521	.474
3	2	1.417	.980	.773	.658	.587	.539
4	2	1.659	1.296	1.090	.958	.868	.805
5	2	1.574	1.193	.984	.824	.757	.704
6	2	1.659	1.309	1.116	1.014	.934	.884
7	2	1.734	1.407	1.214	1.120	1.050	.999
8	2	1.809	1.498	1.356	1.258	1.201	1.162
9	3	1.539	1.069	.844	.702	.613	.547
10	3	1.800	1.401	1.198	1.043	.948	.882
11	3	1.855	1.500	1.276	1.122	1.021	.950
12	3	1.897	1.550	1.333	1.188	1.094	1.026
13	3	2.055	1.723	1.516	1.393	1.290	1.220
14	3	1.971	1.649	1.447	1.312	1.214	1.157
15	3	2.095	1.865	1.726	1.616	1.538	1.488
16	3	2.179	1.940	1.809	1.712	1.639	1.588
17	4	1.430	1.005	.797	.648	.572	.519

those for Dataset 13, in which the amount of variance accounted for by factors 2 and 3 was the same as in Dataset 14, but the factors were of equal strength; there was a trend for the difference between inaccuracies for the two datasets to increase as test length increased, with Dataset 14 resulting in lower mean inaccuracy.

As in the bias data, small effects on inaccuracy and RMSE were observed for the 4-factor ASVAB structure (Dataset 17). Both inaccuracy and RMSE decreased with increasing test length. For a 15-item test, inaccuracy was .581 for Dataset 17 versus .451 for Dataset 1; corresponding RMSE values were .797 and .603.

Although not shown here, similar trends for bias, inaccuracy, and RMSE were observed in the 1.5 and 2.0 OSAF datasets. That is, mean bias became increasingly negative with increasing multidimensionality and test length, whereas mean inaccuracy and RMSE tended to decrease with those variables. In general, however, the magnitudes of the evaluative indices were lower, indicating less effect of multidimensionality with a stronger first factor.

Conditional Indices

Effect of Test Length

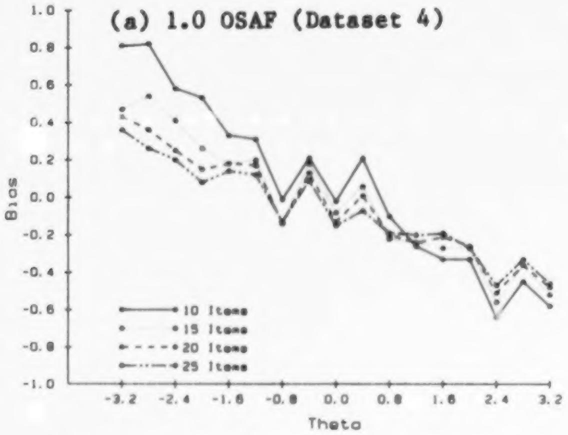
Bias. Figures 1a through 1c show values of mean bias at each of 17 θ levels. Each figure compares the mean bias level across four different test lengths (10, 15, 20, and 25 items) for datasets derived from a 2-factor structure with the second factor 1/3 as strong as the first factor. The first factor for the datasets in Figure 1a was OSAF; in Figure 1b the first factor was 1.5 OSAF; and in Figure 1c it was 2.0 OSAF.

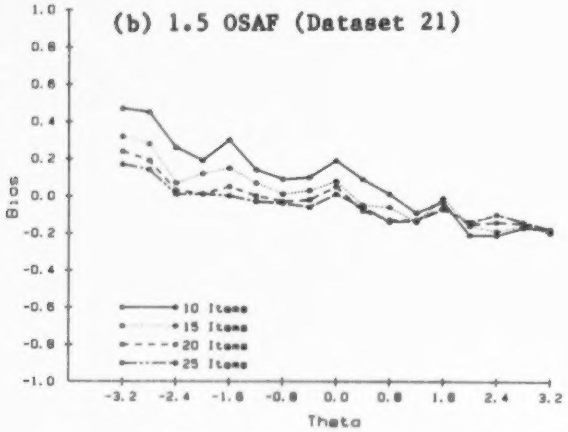
In each of these three figures the mean bias level generally decreases with increasing test length. This pattern is disrupted somewhat between θ levels of -.80 to +.80, where the bias fluctuates around 0.0 and no test length consistently shows a smaller mean bias level. Bias is most variable for the 10-item test length and least variable for the 25-item test. Regardless of the strength of the first factor, the mean bias values at θ levels greater than .80 converge for all four test lengths. Similar patterns of bias across test lengths were observed for the other datasets. In general, bias was negative for θs below the mean and positive for θs above the mean, although this effect was much less pronounced for the 1.5 OSAF datasets (Figure 1b) than for the 1.0 or 2.0 OSAF datasets (Figures 1a and 1c).

Inaccuracy. Figure 2 compares the mean inaccuracy levels at each of four different test lengths (10, 15, 20, and 25 items) across all 17 θ levels for Dataset 29, in which the first factor is 1.5 OSAF, the second factor is 1/2 as strong as the first factor, and the third factor is 1/4 as strong as the first factor. Inaccuracy tended to decrease with increasing test length. Inaccuracy levels for the 10-item test length varied across θ levels and were most constant for the 25-item test. This same pattern held for the comparisons across test length of the mean inaccuracy values for each of the 45 datasets.

RMSE. Comparison of the conditional RMSE values for the same dataset

Figure 1
Conditional Bias of θ Estimates for Tests of 10, 15, 20, and 25 Items
for Datasets with Factor 1 of 1.0, 1.5 and 2.5 OSAF
and Factor 2 One-Third the Strength of Factor 1


(a) 1.0 OSAF (Dataset 4)


(b) 1.5 OSAF (Dataset 21)

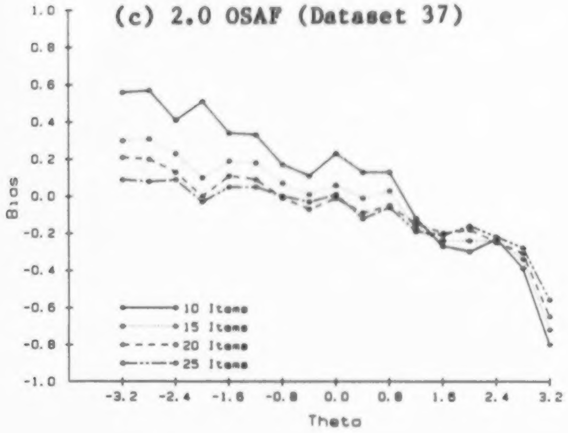

(c) 2.0 OSAF (Dataset 37)

Figure 2
Conditional Inaccuracy of θ Estimates for Tests of 10, 15, 20, and 25 Items for Dataset 29

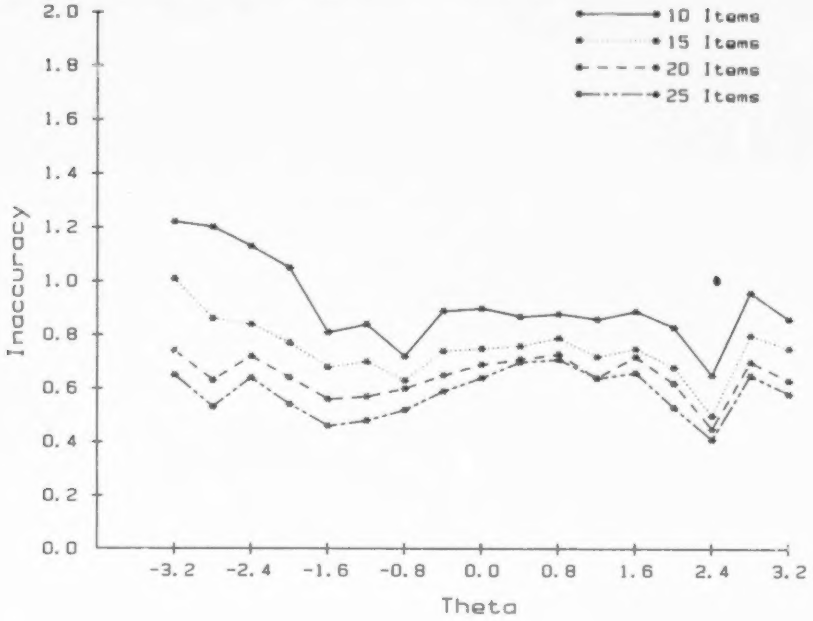
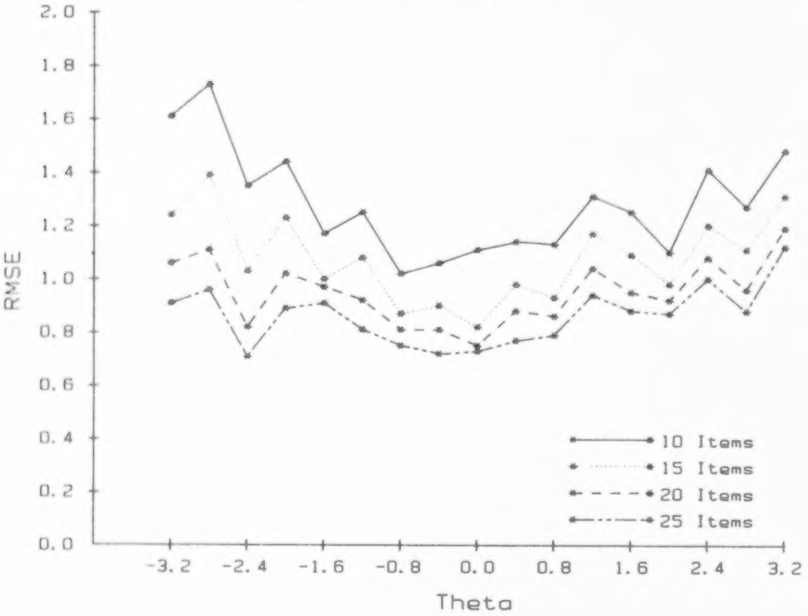


Figure 3
Conditional RMSE of θ Estimates for Tests of 10, 15, 20, and 25 Items for Dataset 4



across different test lengths yielded the same results as for inaccuracy. An example is shown in Figure 3 for Dataset 4. RMSE decreases with increasing test length, and the RMSE values for the shorter test lengths (10 and 15 items) vary more across θ levels than those for the longer tests.

Efficiency. Comparison of the mean efficiency levels for a given dataset across θ levels for a number of test lengths indicated that the efficiency levels increased and followed the same pattern, as test length increased. Figure 4 provides an example of these comparisons for Dataset 29 at 10-, 15-, 20-, and 25-item test lengths.

Since the results for all four conditional indices showed relatively systematic trends as a function of test length, the remainder of the results reported are only for the 15-item test length.

Effect of Multidimensionality

Tables 5 through 8 contain values of the distance measures across 17 θ levels for conditional values of each of the evaluative indices between each UD dataset and each of the MD datasets with the same strength first factor, for tests of 15 items in length. These tables also contain the proportions of the distance measure due to level, scatter, and shape effects.

Bias. Table 5 shows results of the D^2 profile analysis for bias. For the datasets based on OSAF (Datasets 1-17), the UD dataset (Dataset 1) generally had a higher mean bias (.38) and a lower variability (scatter) of bias (2.60) than did the MD datasets (2-17). When a second factor was added to the data (Datasets 2-8), D^2 values tended to increase with increasing strength of the second factor; the exception to this is Dataset 5, in which D^2 values were uniformly lower than in Dataset 4 even though the second factor in Dataset 5 was stronger. The effect proportions show that in all these datasets the vast majority of the differences in bias values as a result of multidimensionality was due to increased scatter; in Datasets 2-8 at least 87% of the differences in bias values from the UD dataset was due to scatter. Level effects accounted for most of the remaining effect for most of these datasets, with the exception of Dataset 2, in which the shape effect was slightly stronger than the level effect.

Similar results were observed for the 2-factor structure in which the first factor was strengthened. For Datasets 19-25, based on 1.5 OSAF, overall D^2 values increased regularly with increasing multidimensionality, but the absolute values of D^2 were smaller than for the 1.0 OSAF data. For Datasets 35-40 a similar but more irregular trend is evident, with smaller values of D^2 than for 1.0 OSAF or 1.5 OSAF, particularly for the higher strength second factors (Datasets 37-40). The effect proportions for these datasets are similar to those for the 1.0 OSAF data, though there is a tendency for multidimensionality to result in slightly greater differences in level, with consequent reductions in the scatter effect.

Figure 5 shows a typical result for bias with increasing multidimensionality for the 1.5 OSAF data. (The values plotted in this figure and in the other figures following are the untransformed values, so that the means and scatters

Figure 4
Conditional Efficiency of θ Estimates for Tests of 10, 15, 20, and 25 Items for Dataset 29

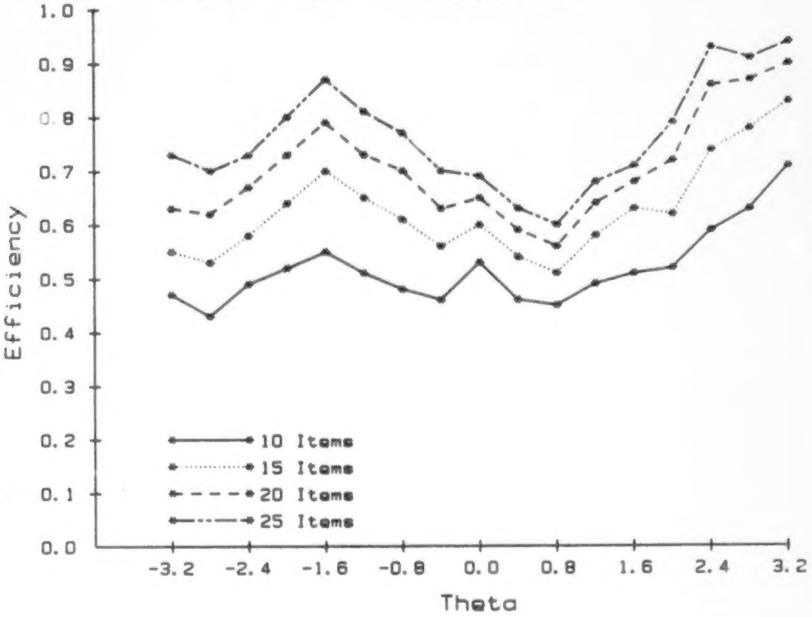


Figure 5
Conditional Bias of θ Estimates for Datasets 18, 21, 23, and 25

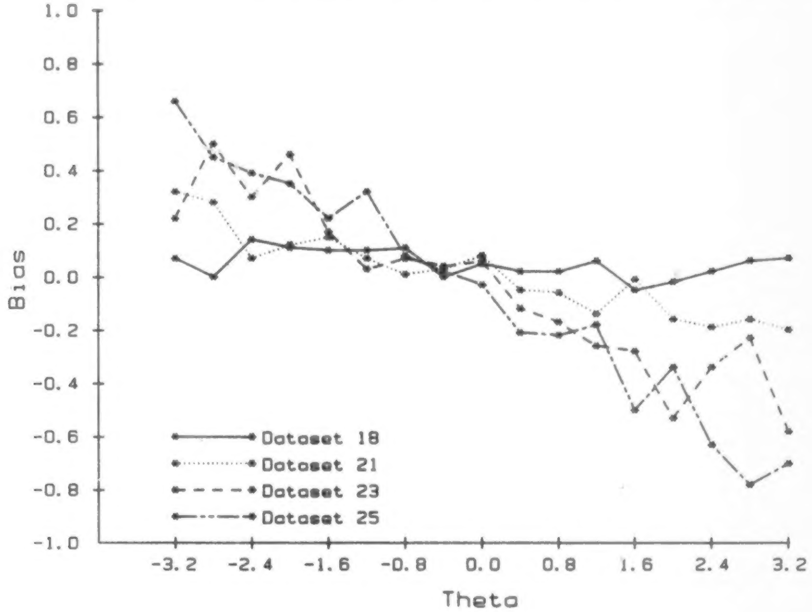


Table 5
Elevation (Mean) and Scatter of Bias (× 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

			Difference Between					Effect Proportion		
Dataset	Mean	Scatter	Means	Scatter	D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Level	Scatter	Shape
1 (UD)	.38	2.60								
2	.31	3.76	.07	-1.17	25.524	25.439	2.468	.003	.900	.097
3	-.03	5.11	.41	-2.51	39.584	36.733	2.295	.072	.870	.058
4	-.20	13.57	.58	-10.98	212.407	206.674	2.445	.027	.962	.012
5	-.14	9.56	.52	-6.96	110.142	105.592	2.304	.041	.938	.021
6	-.22	14.13	.60	-11.53	225.007	218.948	2.345	.027	.963	.010
7	-1.19	18.04	1.57	-15.44	388.634	346.573	2.309	.108	.886	.006
8	-.98	17.97	1.36	-15.37	378.507	346.969	2.371	.083	.910	.006
9	.34	5.01	.04	-2.42	29.444	29.419	1.811	.001	.938	.062
10	-.79	12.59	1.17	-10.00	206.466	183.387	2.552	.112	.876	.012
11	-.37	18.31	.75	-15.72	354.696	345.152	2.063	.027	.967	.006
12	-.43	20.52	.81	-17.93	438.638	427.464	1.990	.025	.970	.005
13	.62	22.51	-.24	-19.92	535.791	534.840	2.365	.002	.994	.004
14	-1.23	22.27	1.60	-19.67	565.182	521.402	2.328	.077	.918	.004
15	-.56	28.55	.94	-25.96	852.904	838.016	2.216	.017	.980	.003
16	1.03	31.53	-.65	-28.94	1044.31	1037.20	2.441	.007	.991	.002
17	.06	3.84	.32	-1.25	22.926	21.212	1.971	.075	.839	.086
18 (UD)	.50	2.09								
19	.41	2.66	.09	-.57	4.011	3.876	.640	.034	.807	.159
20	.20	5.60	.30	-3.52	29.204	27.642	1.306	.054	.902	.045
21	.09	6.23	.41	-4.15	40.071	37.153	1.536	.073	.889	.038
22	-.09	9.04	.60	-6.95	79.711	73.661	1.343	.076	.907	.017
23	-.39	12.62	.89	-10.54	154.682	141.234	1.146	.087	.906	.007
24	-1.03	16.58	1.53	-14.49	296.347	256.390	1.343	.135	.861	.005
25	-.65	17.28	1.16	-15.19	293.890	271.188	1.121	.077	.919	.004
26	.39	5.07	.12	-2.98	22.341	22.108	1.252	.010	.934	.056
27	.38	14.73	.12	-12.64	192.998	192.752	1.073	.001	.993	.006
28	.03	15.61	.47	-13.53	229.319	225.539	1.307	.016	.978	.006
29	-.14	14.84	.64	-12.76	206.511	199.462	1.187	.034	.960	.006
30	.03	17.84	.48	-15.75	301.909	298.055	1.341	.013	.983	.004
31	.50	10.25	.00	-8.17	88.514	88.514	1.019	.000	.988	.012
32	-.28	18.49	.78	-16.40	332.691	322.277	1.379	.031	.965	.004
33	.41	3.89	.09	-1.80	12.352	12.208	1.105	.012	.899	.089
34 (UD)	.76	2.08								
35	.64	5.34	.12	-3.26	23.025	22.777	1.095	.011	.942	.048
36	-.02	8.11	.78	-6.03	55.720	45.257	.528	.188	.803	.009
37	-.28	10.74	1.05	-8.66	108.624	89.974	.671	.172	.822	.006
38	.07	9.65	.70	-7.57	77.603	69.378	.604	.106	.886	.008
39	.29	13.08	.48	-11.00	139.271	134.398	.494	.028	.968	.004
40	.21	11.62	.55	-9.54	109.813	104.641	.562	.047	.948	.005
41	.68	8.38	.08	-6.30	48.956	48.835	.527	.002	.987	.011
42	.46	9.94	.30	-7.86	71.616	70.057	.400	.022	.973	.006
43	.69	13.02	.08	-10.94	138.069	137.966	.677	.001	.994	.005
44	.75	13.94	.01	-11.86	155.258	155.254	.503	.000	.997	.003
45	-1.58	25.93	2.35	-23.85	687.320	593.785	.463	.136	.863	.001

are 1/10 of the comparable values in Tables 5 through 8.) This figure shows the effect of the strength of the second factor increasing from 1/3 of the first factor (Dataset 21) to 2/3 (Dataset 23) to 1.0 (Dataset 25). Bias for the UD dataset (Dataset 18) is close to zero throughout the θ range. For the MD datasets bias is close to zero for θ values close to 0.0, but it increases as the levels progress toward either extreme, resulting in the increased scatter due to increasing multidimensionality. Bias values are generally positive for θ values less than 0.0 and negative for θ values greater than 0.0. For Dataset 21 with the smallest second factor (1/3) bias is not substantially different from the UD dataset, except at extreme θ values; the major effect on bias for these datasets seems to occur for Dataset 23 (factor 2 = 2/3), with the additional 1/3 added to factor 2 in Dataset 25 resulting in generally little additional bias.

Results for the 3-factor datasets (9-16, 26-32, and 41-44) are also in Table 5. For the 1.0 OSAF data, overall D^2 increased regularly with increasing strength of the second and third factors; for the 1.5 OSAF data, values of D^2 were considerably lower, indicating less effect of increased strength of the second and third factors with the stronger first factor; this trend is further supported by Datasets 41-44 (2.0 OSAF), in which overall D^2 values were the lowest for all the 3-factor datasets. For all but one of the 3-factor datasets over 90% of the difference in bias values between the UD and MD datasets was due to scatter (the exception being Dataset 10 with .876), with secondary effects generally attributable to level effects.

Increasing dimensionality from two to three factors while holding constant total proportion of variance accounted for by the factors resulted in increased scatter of bias in most cases. For example, Dataset 6 was a 2-factor structure with the second factor 2/3 of the first, whereas in Dataset 11 both the second and third factors were 1/3 of the first factor. For Dataset 6 overall D^2 was 225, whereas Dataset 11 obtained a D^2 value of 355; in both cases the proportion of D^2 due to scatter was about .96. A similar effect was observed with the 1.5 OSAF data--overall D^2 for Dataset 23 was 155, whereas that for Dataset 28 was 229. The 2.0 OSAF data did not, however, exhibit this effect since overall D^2 for Datasets 39 and 43 were 138 and 139, respectively.

When results from the ASVAB 4-factor structure were compared to those of the relevant UD datasets, very minor effects on bias were observed when OSAF was used (Dataset 17) or when the first factor was increased to 1.5 its original strength (Dataset 33). In both cases mean bias was lower for the ASVAB structure than for the UD structure, though the scatter of the bias was slightly higher. The minor differences in bias for these datasets were, like the other MD structures, primarily due to scatter (.839 for Dataset 17 and .899 for Dataset 33). In contrast to the other MD structures, however, secondary effects were more important for shape than for level, indicating that the ASVAB structure changed the ordering of bias values across the 17 θ levels in comparison to the datasets. However, since there were very small effects on bias due to the ASVAB structure (overall D^2 values of 23 and 12), the shape effects are likely not important.

Using the ASVAB structure with the 2.0 OSAF data (Dataset 45) resulted in the largest overall D^2 for Datasets 35-45, a result considerably different than that observed for Datasets 17 and 33. These data indicate that bias increased substantially both in overall level and variability from the comparable UD dataset, with 86% of the differences in bias due to scatter and 14% due to level. Since factors 2-4 were the same in all three ASVAB datasets, this difference can be attributed only to the increased absolute strength of the first factor in Dataset 45.

Inaccuracy. Table 6 contains the distance measures computed between the inaccuracy profiles of the UD datasets and each of the MD datasets with the same strength first factor. For the 2-factor structures overall, D^2 generally increases with increasing strength of the second factor in both the datasets based on 1.0 OSAF (2-8) and those based on 1.5 OSAF (19-25), with a similar but more irregular trend in the datasets based on 2.0 OSAF. As for the bias criterion, the value of D^2 tends to decrease as the strength of the first factor increases--even though the relative strength of the second factor is the same--indicating less effect on inaccuracy as the strength of the first factor increases. The effect proportions for these data show that differences in inaccuracy values were primarily the result of level effects that tended to increase with increased strength of the second factor. This increasing level effect occurred at the expense primarily of the scatter effect which, with a few exceptions, tended to decrease with increasing strength of the second factor. The only exception to the predominance of the level effect occurred when the second factor was 1/8 as strong as the first factor in Dataset 2, in which case the scatter effect was .547 and the level effect was .382; in the comparable datasets (19 and 35) with similar strength second factors but stronger first factors, the scatter effect was also relatively large. However, in all three of these datasets, D^2 was relatively small, indicating little effect on inaccuracy with a weak second factor.

A similar pattern was observed for the 3-factor structures (Datasets 9-16, 26-32, and 41-44). D^2 tended to increase with increasing strength of the second and third factors, although the trend was more irregular for the 1.5 and 2.0 OSAF data. In all cases level accounted for a minimum of 86% of the squared difference between inaccuracy values for the UD and MD datasets. There was also a marked tendency for the effect of the second and third factors to diminish substantially as the first factor increased in strength. For example, in Dataset 11 based on 1.0 OSAF and second and third factors each 1/3 as strong as the first factor, D^2 was 382 with 96% due to level; in Dataset 28 based on 1.5 OSAF D^2 was 326 with 98% due to level, and in Dataset 43 based on 2.0 OSAF D^2 was 141 with 88% due to level.

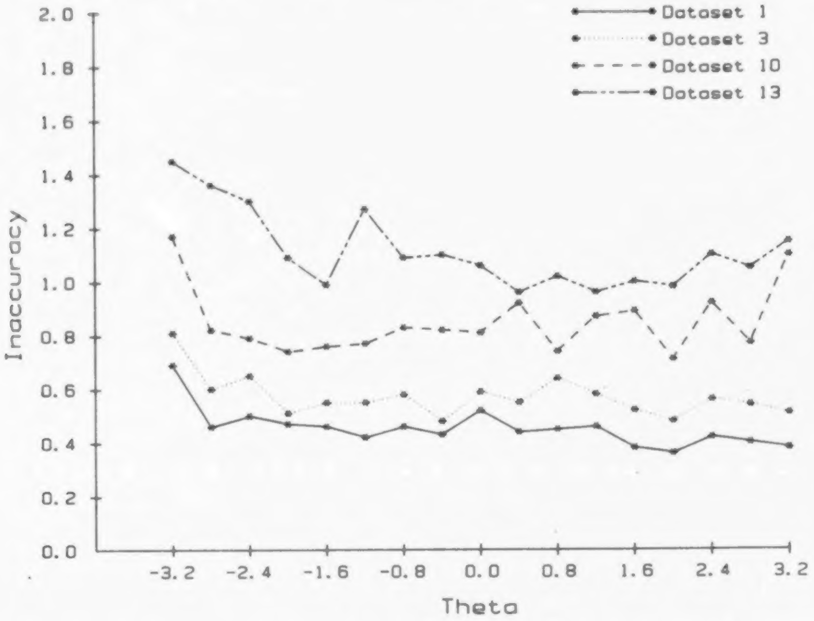
When the number of factors was increased from 2 to 3 while holding constant the proportion of variance accounted for by factors beyond the first, D^2 tended to increase, indicating a greater effect on inaccuracy for a larger number of factors. For example, in the 1.0 OSAF data, D^2 for Dataset 5 (2 factors, second factor 1/2 of first factor) was 130, whereas in Dataset 10 (3 factors, second and third factors each 1/4 of first factor) D^2 was 288; similar effects were observed in the 1.5 OSAF data for Datasets 22 versus 27 (D^2 = 91 vs. 275) and in the 2.0 OSAF data for Datasets 38 and 42 (D^2 = 71 vs. 98). Figure 6 illustrates the typical level effect found for inaccuracy within the 1.0 OSAF data. Dataset 3 with a weak (1/4) second factor results in inaccuracy values close to UD Dataset 1, whereas inaccuracy increases for Dataset 10 with two factors each 1/4 of

Table 6
Elevation (Mean) and Scatter of Inaccuracy (x 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

Dataset	Mean	Scatter	Difference Between Means	Difference Between Scatter	D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Effect Proportion Level	Effect Proportion Scatter	Effect Proportion Shape
1 (UD)	4.52	2.97								
2	5.17	2.95	-.65	.02	18.998	11.735	1.336	.382	.547	.070
3	5.71	3.12	-1.19	-.15	26.202	2.245	.240	.914	.077	.009
4	7.70	3.55	-3.19	-.57	190.049	17.495	1.628	.908	.083	.009
5	7.17	3.84	-2.65	-.86	129.530	10.312	.839	.920	.073	.006
6	7.91	3.19	-3.39	-.21	213.406	17.736	1.868	.917	.074	.009
7	8.61	6.10	-4.09	-3.13	329.511	44.879	1.932	.864	.130	.006
8	9.98	5.11	-5.46	-2.13	539.101	31.407	1.768	.942	.055	.003
9	6.03	2.92	-1.51	.06	45.036	6.235	.719	.862	.122	.016
10	8.48	4.95	-3.96	-1.97	287.725	20.615	1.138	.928	.068	.004
11	9.16	4.67	-4.64	-1.70	381.703	15.607	.916	.959	.038	.002
12	9.48	5.67	-4.96	-2.70	456.103	38.004	1.822	.917	.079	.004
13	11.13	5.87	-6.61	-2.89	766.013	22.639	.818	.970	.028	.001
14	10.62	6.32	-6.10	-3.35	698.322	65.996	2.914	.905	.090	.004
15	12.99	5.29	-8.47	-2.31	1256.73	37.017	2.014	.971	.028	.002
16	13.71	6.01	-9.19	-3.03	1459.01	22.300	.733	.985	.015	.001
17	5.81	2.47	-1.29	.51	36.439	8.198	1.083	.775	.195	.030
18 (UD)	3.47	2.24								
19	3.92	2.25	-.46	-.00	5.919	2.375	.471	.599	.322	.080
20	4.65	2.34	-1.18	-.10	26.112	2.344	.445	.910	.073	.017
21	4.69	2.48	-1.22	-.24	29.550	4.181	.741	.859	.116	.025
22	5.67	3.61	-2.20	-1.36	91.193	8.891	.868	.903	.088	.010
23	6.52	4.84	-3.05	-2.60	176.205	18.184	1.054	.897	.097	.006
24	7.37	6.08	-3.90	-3.84	286.368	27.907	.964	.903	.094	.003
25	8.30	5.40	-4.84	-3.16	424.959	27.536	1.450	.935	.061	.003
26	4.92	3.41	-1.45	-1.17	39.773	3.970	.339	.900	.091	.009
27	7.36	5.70	-3.89	-3.46	275.196	17.818	.458	.935	.063	.002
28	7.79	3.38	-4.32	-1.13	325.967	8.300	.925	.975	.023	.003
29	7.49	4.26	-4.02	-2.02	284.848	9.615	.579	.966	.032	.002
30	8.28	4.64	-4.81	-2.40	413.759	20.685	1.436	.950	.047	.003
31	6.87	3.08	-3.40	-.83	203.233	6.271	.808	.969	.027	.004
32	8.64	3.89	-5.18	-1.65	474.291	18.741	1.836	.960	.036	.004
33	4.35	3.06	-.88	-.82	17.426	4.166	.509	.761	.210	.029
34 (UD)	2.78	3.46								
35	3.59	3.62	-.80	-.16	19.489	8.578	.683	.560	.405	.035
36	4.20	3.76	-1.42	-.30	41.527	7.347	.558	.823	.163	.013
37	4.80	4.32	-2.02	-.86	89.533	20.142	1.298	.775	.210	.014
38	4.67	3.50	-1.88	-.04	70.768	10.465	.864	.852	.136	.012
39	5.70	4.25	-2.92	-.79	158.053	13.192	.855	.917	.078	.005
40	5.42	4.70	-2.64	-1.24	127.626	9.232	.474	.928	.069	.004
41	4.73	5.21	-1.95	-1.75	74.287	9.840	.377	.868	.127	.005
42	5.11	3.10	-2.33	.36	97.648	5.296	.482	.946	.049	.005
43	5.48	6.02	-2.70	-2.56	140.984	17.415	.522	.876	.120	.004
44	5.91	5.85	-3.13	-2.39	180.958	14.843	.450	.918	.080	.002
45	8.94	9.07	-6.15	-5.61	736.603	92.771	1.953	.874	.123	.003

the first factor, and increases again in Dataset 13 as factors 2 and 3 are again increased to 1/2 of the first factor.

Figure 6
Conditional Inaccuracy of θ Estimates for Datasets 1, 3, 10, and 13



The ASVAB factor structure (Datasets 17, 33, and 45) had slightly greater effects on overall D^2 for inaccuracy (Table 6) than it did for bias (Table 5). Similar to the bias data, however, the ASVAB structure resulted in lowest D^2 for the 1.5 OSAF data (Dataset 33) and a very high value of D^2 in the 2.0 OSAF data (Dataset 45). For all three datasets D^2 was primarily attributable to differences in level of conditional inaccuracy, with a secondary effect due to scatter of the inaccuracy values.

RMSE. The results for RMSE, shown in Table 7, have some similarity to those for inaccuracy. That is, for both the 2- and 3-factor structures D^2 generally increased as the strength of factors beyond the first increased. In addition, the magnitude of D^2 decreased with increasing strength of the first factor, indicating that the effect of factors beyond the first factor on RMSE was less with a stronger first factor, even though succeeding factors were proportionally as strong. In contrast to the inaccuracy results, however, for the 2-factor structures (Datasets 2-8, 19-25, 35-40), MD datasets resulted in RMSE values that were more variable than the UD datasets, as indicated by D^2 scatter proportions in the range of .10 to .20 for most of the 1.0 and 1.5 OSAF structures, and above .20 for many of the 2.0 OSAF datasets (35 to 38). With only one exception (Dataset 2), however, the predominant effect of multidimensionality was to increase the level of RMSE in all datasets, with the greatest level effects observed in the 1.5 OSAF data.

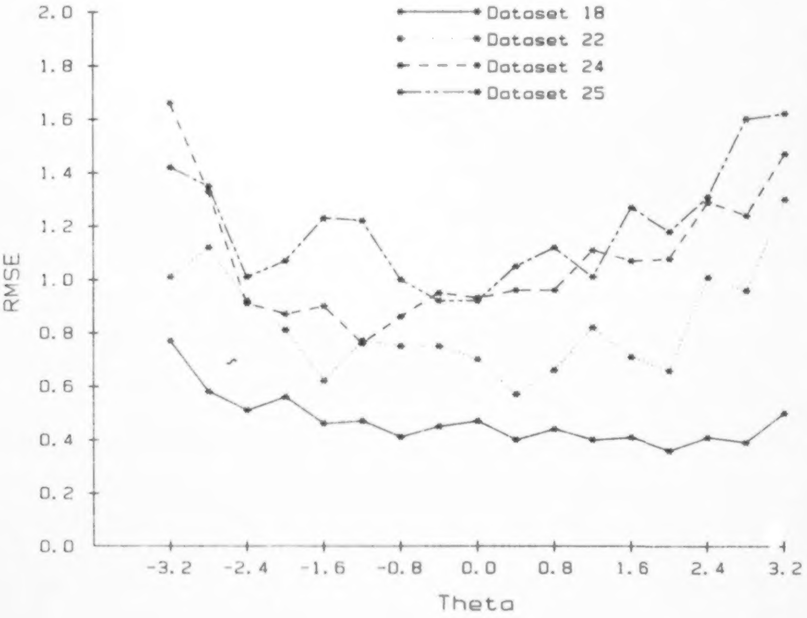
Table 7
Elevation (Mean) and Scatter of RMSE (× 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

			Difference Between					Effect Proportion		
Dataset	Mean	Scatter	Means	Scatter	D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Level	Scatter	Shape
1 (UD)	5.92	5.04								
2	6.85	4.74	-.93	.29	52.258	37.554	1.568	.281	.689	.030
3	7.63	5.06	-1.71	-.02	57.435	7.568	.297	.868	.127	.005
4	10.79	6.43	-4.87	-1.39	460.096	56.266	1.677	.878	.119	.004
5	9.74	6.68	-3.82	-1.64	284.922	36.997	1.020	.870	.126	.004
6	11.05	6.57	-5.13	-1.53	508.037	60.475	1.755	.881	.116	.003
7	11.92	9.60	-6.00	-4.56	728.676	116.171	1.972	.841	.157	.003
8	13.43	8.02	-7.51	-2.98	1038.69	78.628	1.727	.924	.074	.002
9	8.36	5.07	-2.44	-.03	117.990	16.686	.653	.859	.136	.006
10	11.82	8.00	-5.91	-2.96	642.629	49.761	1.018	.923	.076	.002
11	12.64	7.41	-6.72	-2.38	811.771	43.702	1.019	.946	.053	.001
12	13.12	9.74	-7.21	-4.71	994.308	111.601	1.822	.888	.110	.002
13	15.02	8.76	-9.10	-3.72	1467.72	59.222	1.029	.960	.040	.001
14	14.29	9.25	-8.37	-4.21	1326.61	135.073	2.518	.898	.100	.002
15	17.16	8.18	-11.24	-3.14	2229.70	80.075	1.704	.964	.035	.001
16	17.95	9.31	-12.03	-4.27	2525.26	63.754	.971	.975	.025	.000
17	7.90	4.31	-1.98	.72	86.201	19.562	.876	.773	.217	.010
18 (UD)	4.70	3.92								
19	5.39	4.53	-.69	-.61	13.904	5.817	.306	.582	.396	.022
20	6.79	5.29	-2.08	-1.37	84.610	10.949	.438	.871	.124	.005
21	6.84	5.32	-2.14	-1.41	97.021	19.364	.834	.800	.191	.009
22	8.32	7.85	-3.62	-3.93	271.126	48.945	1.089	.819	.177	.004
23	9.68	8.32	-4.98	-4.41	479.630	58.372	1.195	.878	.119	.002
24	10.80	9.70	-6.10	-5.78	702.920	71.289	.995	.899	.100	.001
25	11.94	8.62	-7.23	-4.70	962.968	73.388	1.521	.924	.075	.002
26	7.15	6.21	-2.44	-2.29	112.695	11.181	.244	.901	.097	.002
27	10.80	9.14	-6.09	-5.22	678.314	46.987	.551	.931	.068	.001
28	11.28	6.41	-6.58	-2.50	770.580	34.305	1.118	.955	.043	.001
29	10.99	6.89	-6.28	-2.97	694.135	22.728	.515	.967	.032	.001
30	12.21	8.97	-7.50	-5.05	1025.00	68.198	1.215	.933	.065	.001
31	9.97	5.31	-5.27	-1.39	487.982	15.980	.676	.967	.031	.001
32	12.65	7.08	-7.95	-3.16	1122.39	49.268	1.417	.956	.043	.001
33	6.29	6.25	-1.59	-2.33	55.572	12.693	.297	.772	.223	.005
34 (UD)	4.09	6.65								
35	5.65	7.47	-1.56	-.81	78.437	37.022	.732	.528	.463	.009
36	6.70	7.21	-2.62	-.56	163.897	47.642	.987	.709	.285	.006
37	7.79	8.43	-3.70	-1.78	322.126	89.233	1.534	.723	.272	.005
38	7.61	8.06	-3.52	-1.41	276.660	65.480	1.184	.763	.232	.004
39	9.07	8.79	-4.98	-2.14	485.606	64.437	1.024	.867	.131	.002
40	8.68	8.70	-4.59	-2.04	405.497	47.229	.744	.884	.115	.002
41	7.75	9.10	-3.66	-2.44	264.622	37.109	.515	.860	.138	.002
42	8.32	6.12	-4.23	.53	323.970	20.145	.488	.938	.061	.002
43	8.97	10.26	-4.89	-3.61	462.204	56.519	.638	.878	.121	.001
44	9.48	10.08	-5.39	-3.43	546.413	53.325	.620	.902	.096	.001
45	13.12	12.92	-9.03	-6.27	1585.95	200.102	1.871	.874	.125	.001

Figure 7 shows a typical example of the RMSE results. This figure displays RMSE values for the 1.5 OSAF UD dataset (18) and MD Datasets 22, 24, and 25, in which the strength of the second factor increased respectively from 1/2 to 3/4 to 1.0 of the first factor. As can be seen, values of RMSE increased with increasing strength of the second factor, with only minor changes in their variability.

Figure 7

Conditional RMSE of θ Estimates for Datasets 18, 22, 23, and 25



The patterns of RMSE results for the ASVAB data structures were similar to those for inaccuracy. Lowest D^2 was observed for the 1.5 OSAF data (Dataset 33), whereas highest occurred for 2.0 OSAF. Even though the ASVAB structure included four factors, D^2 values for the 1.0 and 1.5 OSAF structures were in the range of those observed for 2-factor structures with second factors 1/8 to 1/4 those of the first factor (e.g., Datasets 2, 3, 19, 20). The ASVAB structure tended to result in D^2 values with a higher scatter effect for the 1.0 and 1.5 OSAF datasets, in comparison to most of the other MD datasets, indicating more variability in RMSE values as a function of θ levels than was evident in the corresponding UD datasets.

Efficiency. D^2 values for efficiency are in Table 8. With the exception of Dataset 2, the predominant difference in efficiency between the MD and UD datasets in the 2-factor data for 1.0 OSAF (Datasets 2-8) and 1.5 OSAF (Datasets 2-8 and Datasets 19-25) was due to level; MD structures resulted in fairly constant levels of lower efficiency in comparison to UD structures. In the 1.0 OSAF datasets the scatter/variability of observed efficiency values tended to

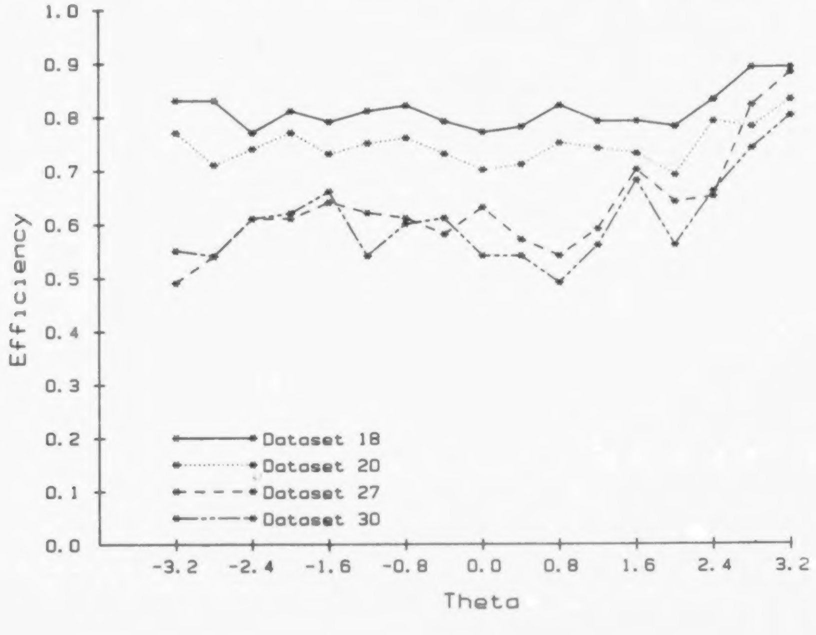
Table 8
Elevation (Mean) and Scatter of Efficiency (× 10) for Unidimensional (UD) and Multidimensional Datasets, Differences Between Elevation and Scatter, Total D^2 Index, D^2 with Elevation Removed ($D^{2\prime}$), D^2 with Elevation and Scatter Removed ($D^{2\prime\prime}$), and Proportion of D^2 Due to Level, Scatter, and Shape, for Tests of 15 Items

			Difference Between					Effect Proportion		
Dataset	Mean	Scatter	Means	Scatter	D^2	$D^{2\prime}$	$D^{2\prime\prime}$	Level	Scatter	Shape
1 (UD)	8.18	1.96								
2	7.92	1.98	.26	-.02	2.620	1.481	.383	.435	.419	.146
3	7.75	1.63	.42	.33	4.080	1.031	.289	.747	.182	.071
4	7.08	1.33	1.09	.63	22.520	2.169	.682	.904	.066	.030
5	7.31	1.37	.86	.59	13.890	1.179	.311	.915	.062	.022
6	7.05	1.16	1.12	.80	23.450	1.991	.597	.915	.059	.025
7	6.79	1.56	1.39	.40	37.060	4.298	1.355	.884	.079	.037
8	6.65	1.32	1.53	.64	42.840	3.075	1.033	.928	.048	.024
9	7.53	2.03	.65	-.08	8.060	.942	.235	.883	.088	.029
10	6.74	2.22	1.44	-.27	37.140	2.119	.471	.943	.044	.013
11	6.51	2.07	1.67	-.11	48.680	1.235	.302	.975	.019	.006
12	6.46	2.42	1.71	-.47	53.550	3.738	.742	.930	.056	.014
13	6.15	2.24	2.03	-.28	72.690	2.675	.593	.963	.029	.008
14	6.27	1.84	1.91	.11	68.460	6.709	1.857	.902	.071	.027
15	5.91	1.08	2.27	.88	89.980	2.335	.741	.974	.018	.008
16	5.64	2.12	2.54	-.16	113.460	3.681	.880	.968	.025	.008
17	7.71	1.55	.47	.40	4.680	.915	.247	.804	.143	.053
18 (UD)	8.11	1.44								
19	7.92	1.28	.19	.16	1.250	.609	.316	.512	.235	.253
20	7.46	1.42	.65	.02	8.110	.862	.421	.894	.054	.052
21	7.52	1.56	.59	-.12	7.010	1.009	.441	.856	.081	.063
22	7.15	1.58	.96	-.14	17.860	2.039	.886	.886	.065	.050
23	6.81	2.12	1.30	-.67	31.390	2.660	.723	.915	.062	.023
24	6.45	3.01	1.66	-1.57	55.270	8.159	1.317	.852	.124	.024
25	6.28	1.32	1.84	.12	60.100	2.839	1.481	.953	.023	.025
26	7.35	1.79	.76	-.35	11.390	1.601	.571	.859	.090	.050
27	6.31	3.86	1.81	-2.42	65.930	10.489	.835	.841	.146	.013
28	6.14	3.06	1.98	-1.62	72.320	5.911	.747	.918	.071	.010
29	6.26	3.61	1.85	-2.16	66.580	8.582	.750	.871	.118	.011
30	6.06	3.25	2.05	-1.81	78.710	7.062	.813	.910	.079	.010
31	6.66	2.29	1.45	-.85	37.580	1.982	.383	.947	.043	.010
32	6.18	1.92	1.94	-.48	65.990	2.319	.754	.965	.024	.011
33	7.62	1.81	.49	-.37	5.690	1.638	.573	.712	.187	.101
34 (UD)	7.80	2.06								
35	7.29	6.49	.51	-4.42	39.220	34.869	1.143	.111	.860	.029
36	7.04	6.15	.76	-4.08	43.320	33.379	1.318	.229	.740	.030
37	6.82	5.74	.98	-3.68	47.000	30.791	1.458	.345	.624	.031
38	6.90	6.49	.90	-4.42	50.650	36.880	1.294	.272	.703	.026
39	6.46	5.56	1.34	-3.50	57.280	26.701	1.258	.534	.444	.022
40	6.59	6.17	1.21	-4.11	58.540	33.578	1.311	.426	.551	.022
41	6.76	6.57	1.04	-4.50	53.820	35.599	1.130	.339	.640	.021
42	6.66	6.49	1.14	-4.43	56.380	34.241	1.091	.393	.588	.019
43	6.61	6.79	1.19	-4.72	63.490	39.249	1.210	.382	.599	.019
44	6.42	6.65	1.38	-4.58	69.750	37.265	1.184	.466	.517	.017
45	5.46	4.89	2.34	-2.83	122.740	29.561	2.135	.759	.223	.017

decrease with increasing strength of the second factor, with a somewhat more irregular trend observed for the comparable 1.5 OSAF datasets. For the 3-factor structures in the 1.0 and 1.5 OSAF datasets (Datasets 9-16 and 26-32), the predominant result was an overall reduction of efficiency values as the strength of the second and third factors increased. The level effect for these datasets tended to be in the high .80s and low .90s with a minor secondary effect due to scatter. In both the 1.0 and 1.5 OSAF structures, an increase from 2 to 3 factors while maintaining the same proportion of variance in the factors beyond the first led to decreases in efficiency, as shown by D^2 values of 23 for Dataset 6 (second factor 2/3 of the first) and 44 for Dataset 11 (second and third factors each 1/3 of the first).

Figure 8 shows the typical pattern of results for the 1.0 and 1.5 OSAF data. The UD data structure (Dataset 18) shows a fairly flat and high pattern of efficiency with a mean of .811. When a second factor 1/4 the strength of the first factor is added in Dataset 20, mean efficiency drops to .75 with little change in variability or shape. Datasets 27 and 30 show strong effects on efficiency through most of the θ range when two factors are added to the first. However, the strength of the second and third factors seems to have little effect on efficiency since factors 2 and 3 in Dataset 27 were each 1/4 of the first factor, whereas these factors each accounted for 1/2 the variance of the first factor in Dataset 30. The trend observed in Figure 8 for Datasets 27 and 30 appeared for most of the efficiency data--there was a tendency for strong second and third factor structures to have a greater effect for lower θ levels than for higher θ levels. This asymmetry was not evident in the bias, inaccuracy, or RMSE results.

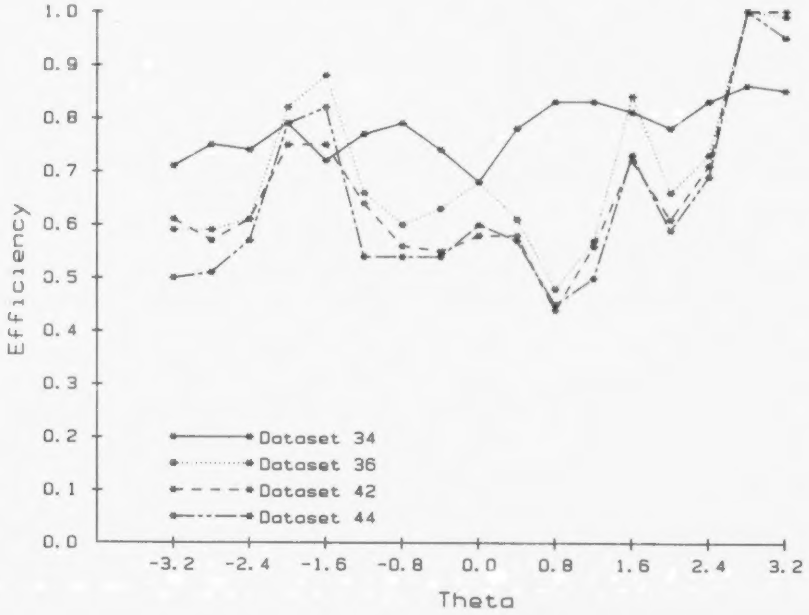
Figure 8
Conditional Efficiency of θ Estimates for Datasets 18, 20, 27, and 30



A different pattern of results emerged for the 2.0 OSAF data. For the UD Dataset 34, mean efficiency (.78) was slightly lower and its scatter higher than for UD Datasets 1 and 18. As the strength of the second and third factors increased, overall D^2 values increased to about the same levels as those observed in comparable 1.0 and 1.5 OSAF data, indicating similar overall reductions in efficiency; for example, D^2 in Dataset 40 (with a second factor 3/4 of the first factor) was 59, whereas the same structure in the 1.5 OSAF data (Dataset 24) resulted in a D^2 of 55. The difference in the 2.0 OSAF data versus the 1.5 and 1.0 OSAF was in the pattern of the efficiency results. Whereas in the latter data structures the predominant D^2 effect was for level, in the 2.0 OSAF data the majority of the change in efficiency due to multidimensionality was due to scatter, with proportions ranging from .86 for Dataset 35 to .44 for Dataset 39.

Figure 9 displays the typical pattern of results for the 2.0 OSAF data structures. UD Dataset 34 has the flattest and generally highest efficiency levels of the datasets plotted. The remainder of the datasets resulted in similar patterns of highly variable efficiency values, all following a similar pattern and differing little, even though Dataset 36 had only two factors with the second factor only 1/4 the strength of the first, whereas Datasets 42 and 44 were 3-factor structures with the second and third factors combined accounting for 1/2 and 3/4 the variance of the first factor, respectively. For all three of these datasets, efficiency values for the MD structures exceeded those of the UD structure for θ values in the range of -1.6 to -2.0 and above about 2.8.

Figure 9
Conditional Efficiency of θ Estimates for Datasets 34, 36, 42, and 44



Results for the ASVAB structures show small reductions in mean efficiency from .82 to .77 in the 1.0 OSAF data (Datasets 1 vs. 17), with a reduction in scatter; a similar small mean effect in the 1.5 OSAF data (.81 vs. .76); and a slight increase in scatter for Datasets 18 versus 33. When the first factor was increased to twice its original strength, addition of the three ASVAB factors resulted in a substantial decrease in mean efficiency and in a substantial increase in the variability of efficiency values; in Dataset 45 (the ASVAB structure) mean efficiency was .55 with scatter of .49, in comparison to values of .78 and .21 for the UD 2.0 OSAF structure (Dataset 34). However, for all three comparisons, level effects accounted for more than 70% of the differences between conditional efficiency levels for the ASVAB data and the comparable UD datasets.

CONCLUSIONS

As the overall degree of multidimensionality (as measured by the sum of the eigenvalues for each factor) in the generated item responses increased, the estimated θ values at each of the seventeen θ levels evaluated deviated further from the true (first factor) θ values. This effect was evident in the comparisons of overall bias, inaccuracy, and root mean square (RMSE) values for datasets with differing degrees of multidimensionality, and in all the conditional indices. These comparisons showed increasing levels of each of these evaluative indices as the multidimensionality of the underlying factor structure increased. The effect was also evident in the decreased efficiencies of datasets when compared to datasets with underlying factor structures that were more unidimensional. Individual θ estimates also ordered individuals differently from the true values, as reflected in the fidelity correlations. The pattern of results, therefore, suggests that maximum information adaptive testing is sensitive to changes in the dimensionality of the responses.

While all degrees of multidimensionality had effects on all the evaluative indices, effects were generally a function of the number of items administered. Thus, for the overall indices in all multidimensional datasets, fidelities increased with increasing test length, and inaccuracy and RMSE decreased, while overall bias tended to change from fairly high positive values for short test lengths to low negative values for the majority of multidimensional structures. For the conditional indices, very similar patterns of results were observed for different test lengths, with level effects (as opposed to scatter or shape effects) predominant for all but the bias index. Even for conditional bias, however, test length effects were roughly proportional for a given θ level. Consequently, while maximum information adaptive testing is affected by deviations from unidimensionality, the data suggest that in many cases, at least for relatively small degrees of multidimensionality, the effects of multidimensionality can be overcome simply by increasing test length. For example, the ASVAB factor structure resulted in a fidelity of .802 for a 15-item test compared to .872 for the UD case. When the multidimensional ASVAB structure was increased to 25 items in length, the fidelity of .871 was essentially the same as that of the 15-item unidimensional test. The same pattern was observed when the first factor of the ASVAB structure was strengthened by 50%.

The overall indices showed, in general, that increasing test length to twice the length of the multidimensional tests will overcome the effects of multidimensionality for multidimensional structures with one or two factors beyond the first that account for up to one-fourth the variance of the first factor. This finding held regardless of the strength of the first factor. Since a similar result was observed for the ASVAB structure (in which factors 2, 3, and 4 accounted for 22%, 13%, and 13% of the first factor, respectively) in the 1.0 and 1.5 OSAF data, the results suggest that the effects on maximum information adaptive testing of multidimensional factor structures in which up to one-third of the variance of the first factor appears in second and third factors, can be overcome by doubling adaptive test length. For degrees of multidimensionality beyond these levels, however, adaptive test lengths would need to be increased well beyond double to overcome the effects of multidimensionality. This conclusion must be qualified, however, when bias of the θ estimates is of concern, since the degree of bias differed at different θ levels.

There was some evidence to suggest that the number of factors (2 vs. 3), and not simply the overall strength of the underlying factor structure, affected θ estimates. For example, a single factor beyond the first had less effect on fidelity than did two factors that accounted for the same amount of variance. In addition, there was more scatter of conditional bias with three factors than with two, even though the proportion of variance in the second and third factors was equal in the two structures. Thus, the more complex factor structures seemed to affect the θ estimates more than the simpler structures. This finding, however, did not appear to extend to the 4-factor ASVAB structures.

Several factors affect the generality of the conclusions drawn from this research. First, the results are limited to the particular multidimensional model used to generate the multidimensional response vectors. Use of other models, such as those reviewed by Reckase and McKinley (1985), may yield different results. The results are also limited to maximum information adaptive testing with maximum likelihood scoring. Third, different factor structures might result in different findings, since only one basic first factor was used in this study. Thus, the study should be replicated varying these factors to further evaluate the robustness of adaptive testing to deviations from the unidimensional item response theory model used to select and to score test items.

REFERENCES

Betz, N. E., & Weiss, D. J. An empirical study of computer-administered two-stage ability testing (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1973.

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1974.

Betz, N. E., & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing (Research Report 75-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1975.

Crichton, L. I. Effects of error in item parameter estimates on adaptive testing. Unpublished doctoral dissertation, University of Minnesota, 1981.

Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Drasgow, F., & Parsons, C. K. Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 1983, 7, 189-199. (Also this volume, pp. 218-232.)

Johnson, M. F., & Weiss, D. J. Parallel forms reliability and measurement accuracy comparison of adaptive and conventional testing strategies. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Kiely, G. L., Zara, A. R., & Weiss, D. J. Alternate forms reliability and concurrent validity of adaptive and conventional tests with military recruits (Draft Final Report of Contract N00123-79-C-1273, submitted to Navy Personnel Research and Development Center). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, January 1983.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing (Research Report 74-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, July 1974.

Larkin, K. C., & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing (Research Report 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1975.

Mattson, J.D. Effects of item parameter error and other factors on trait estimation in latent trait-based adaptive testing. Unpublished doctoral dissertation, University of Minnesota, 1983.

McBride, J. R. Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1977, 1, 121-140.

McBride, J. R., & Martin, J. T. Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press, 1983.

Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. Applied Psychological Measurement, 1984, 8, 155-163.

Reckase, M. D., & McKinley, R. L. Some latent trait theory in a multidimensional space. In D. J. Weiss (Ed.), Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference. Minnepolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1985, pp. 151-177.

Sympson, J. B., Weiss D. J., Ree, M. J. Predictive validity of conventional and adaptive tests in an Air Force training environment (AFHRL-TR-81-40), Brooks Air Force Base TX: Air Force Systems Command, Manpower and Personnel Division, August 1982.

Urry, V. W. Computer-assisted testing: The calibration and evaluation of the verbal ability bank (Technical Study 74-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974.

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975. (a)

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975. (b)

Vale, C. D. Problem: Strategies of branching through an item pool. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975.

Weiss, D. J., & McBride, J. R. Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 1984, 8, 273-285.

ACKNOWLEDGMENT

This research was supported by Contract N00014-79-C-0172, NR 150-433, with the Office of Naval Research, with additional funding from the Air Force Human Resources Laboratory, Army Research Institute, and the Air Force Office of Scientific Research.

USE OF SEQUENTIAL TESTING TO PRESCREEN PROSPECTIVE ENTRANTS INTO MILITARY SERVICE

R. A. WEITZMAN
NAVAL POSTGRADUATE SCHOOL

The objective of this research was to study the possible recruiting station use of a form of sequential testing called selective testing to prescreen applicants for military enlistment.

In selective testing (described by Weitzman, 1982), an applicant responds at a computer terminal to one item at a time until the totality of his/her responses indicates either an acceptance or a rejection decision with preset error probabilities: α, the probability of accepting an applicant who will fail, and β, the probability of rejecting an applicant who would succeed if accepted. Although different applicants generally respond to different numbers of items, the average of these numbers tends to be small (less than 20), primarily depending on the magnitudes of the preset error probabilities. The validity of the selection decision requires that successive items be uncorrelated for applicants who have equal values of the performance variable (the criterion) that the test is used to predict. This local independence requirement was evidently met in a previous application of the method involving 960 Navy enlisted men who had taken both an entrance and a final examination for a technical training course. Application of the method to the entrance examination to predict passing or failing on the final examination resulted in observed error proportions that closely matched preset error probabilities (Weitzman, 1982).

The first use of sequential testing to classify individuals was an application to dichtomous classification by Linn, Rock, and Cleary (1972) of a sequential procedure developed by Armitage (1950) for polychotomous classification. This procedure used the value of an objective function to determine when testing should be terminated. Though related monotonically to this value, the observed rates of classification errors were not subject to control by the procedure.

Selective testing, a form of sequential testing, can both concentrate its accuracy at the cutting score and control the probabilities of selection errors. Selective testing is an adaptation of the sequential probability ratio test (SPRT) developed by Wald (1945). Other testing adaptations of the SPRT apply specifically to the determination of subject matter mastery (Epstein & Knerr, 1978; Ferguson, 1970; Kalisch, 1980; Kingsbury & Weiss, 1980; Reckase, 1980). The mastery decision in each of these adaptations tends to have error rates that are no higher than preset values only for students whose subject matter mastery corresponds to proportions that fall outside an indifference region that the test user must specify.

Selective testing, by contrast, works to control the error rates for everyone. This control requires monitoring a probability-ratio test statistic, computed after each item response, to determine whether the statistic has reached a value farther from one than an upper or lower critical value. Testing continues until the test statistic has reached one or the other of these two critical values.

Before testing, a standardization group of applicants is divided into K quantile groups on the criterion. The test statistic is a function of the proportion (p_{ik}) of standardization group applicants within criterion quantile group *k* who answer item i (i = 1, 2, ..., n) correctly:

$$L_n = \frac{(K - K^* + 1)^{-1}\sum_{k=K^*}^{K}\prod_{i=1}^{n} p_{ik}^{\;x_i}(1-p_{ik})^{1-x_i}}{(K^* - 1)^{-1}\sum_{k=1}^{K^*-1}\prod_{i=1}^{n} p_{ik}^{\;x_i}(1-p_{ik})^{1-x_i}}, \qquad [1]$$

where K* designates the quantile group immediately above the criterion measurement separating success from failure and x_i equals 1 for a correct and 0 for an incorrect response to item *i*. According to Wald (1945), the critical values for L_n are $(1 - \beta)/\alpha$ for an acceptance and $\beta/(1 - \alpha)$ for a rejection decision.

Method

The data consisted of the responses (correct/incorrect) of 1,020 Navy recruits to 200 items of the Armed Services Vocational Aptitude Battery (ASVAB) together with the scores of these recruits on the Armed Forces Qualification Test (AFQT), which functioned as the criterion. The AFQT is actually a composite of four of the ASVAB components: Arithmetic Reasoning (AR), Paragraph Comprehension (PC), Word Knowledge (WK), and Numerical Operations (NO). The 200 ASVAB items used as predictors represented all these components except NO. Table 1 shows the correlations among the AFQT and the eight components of the ASVAB represented by the items used here. In addition to AR, PC, and WK, these components were General Science (GS), Automotive-Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI).

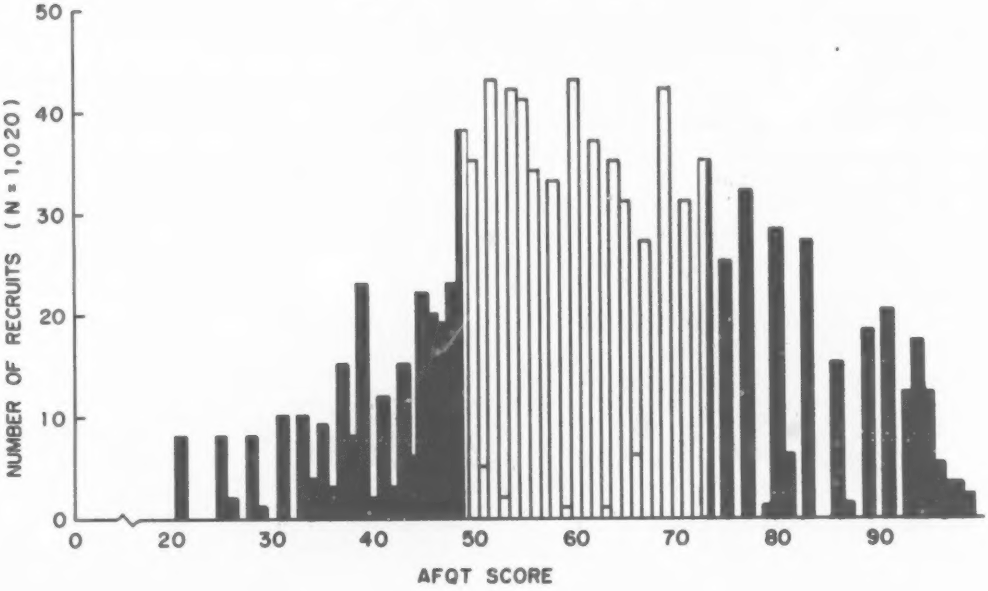
The goal of this research was to predict from a recruit's responses to the ASVAB items whether the recruit would pass the AFQT. The entire group of 1,020 recruits functioned as both the standardization and the applicant group. The histogram in Figure 1 describes the frequency distribution of the AFQT scores for these recruits. The shaded and blank areas represent the different failure rates used--25% and 75%--with the frequency distribution of AFQT scores divided into quartiles for the determination of the p_{ik} values required to compute the test statistic. Use of the overlap treatments described by Weitzman (1982) resolved the problems arising from the overlap apparent in the two boundary score groups.

Table 1
Correlations among AFQT and ASVAB Tests (N=1,032)

	Test								
Test	AFQT	GS	AR	WK	PC	AS	MK	MC	EI
AFQT		.67	.69	.66	.52	.51	.63	.59	.58
GS			.54	.71	.54	.59	.54	.58	.67
AR				.54	.55	.44	.72	.55	.52
WK					.64	.49	.48	.49	.64
PC						.44	.49	.50	.53
AS							.37	.65	.65
MK								.52	.49
MC									.61
EI									

Each failure rate was used in each of three studies exemplifying three methods of item selection. Although every recruit took the entire 200-item test battery, computer runs simulated the sequential procedure by selecting one item at a time. In two of the three studies the order of item selection corresponded directly to the ranking of the correlations between item responses and AFQT scores. In the first study the correlation was a point-biserial coefficient (Method 1); in the second, it was a phi coefficient, with AFQT scores dichotomized at the failure-rate centiles to maximize item discriminability (Method 2).

Figure 1
Frequency Distribution of AFQT Scores
for 1,020 Navy Enlisted Men Showing Failure Rates of
25% (Left Solid) and 75% (Complement of Right Solid)



Selective testing assumes local independence on the AFQT. To select items that most nearly met this assumption, the third study used as an objective function for each candidate item the ratio of the largest partial correlation between the candidate item and each item already selected, controlling for the AFQT, to the point-biserial coefficient used in Method 1. The candidate item selected was the one for which this ratio was smallest (Method 3).

The original intention in all three studies was to truncate the test at item 75. A problem arose that tended to reduce this number, however. This problem was the occurrence of p_{ik} values equal to zero or one that prevented the calculation of the test statistic for two of the 75 items selected for use in each study. Elimination of these two items thus resulted in the truncation item number actually used: 73 in all three studies.

Results

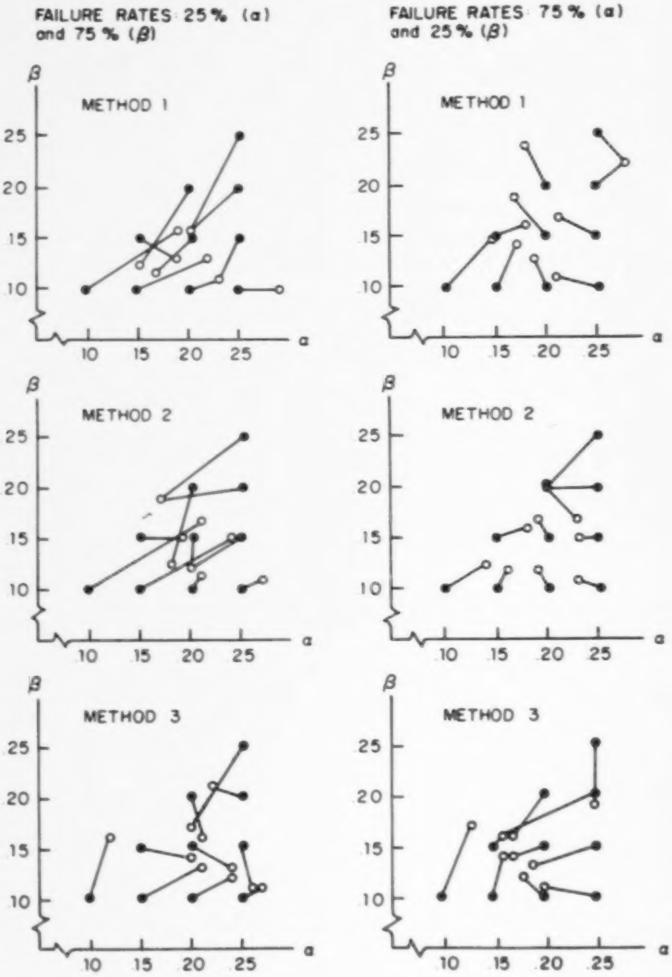
Figure 2 shows the error-rate results. The two graphs in each row compare the expected (solid circles) and observed (open circles) error rates. The three rows represent the three methods of item selection. In each comparison the failure rates differed for the computation of the observed α and β values so that the groups used to compute the α and β coordinates of the open circles consisted of 255 recruits in the left graphs and 765 recruits in the right graphs. The accuracy and stability of the observed α and β values depend on the sizes of the groups used to compute them. The difference in accuracy and stability between the left and right graphs reflects this dependence. The closeness of the observed to the expected values in the right graphs are due largely, if not entirely, to sampling error. Of the three methods, the accuracy appears best for Method 2.

Table 2 presents the mean test lengths (left cell entries) and 73-item frequencies (right cell entries) obtained in all three studies. For Method 2 the means ranged from 3 for $\alpha = \beta = .25$ with the 75% failure rate to 10 for $\alpha = \beta = .10$ with the 25% failure rate. The mean test lengths tended to be well below 73 (the maximum test length), and in no case did more than 19 of the 1,020 recruits require as many as 73 items for a selection decision. For Method 1 and Method 3 both the means and the 73-item frequencies tended to be larger than for Method 2.

Sequential tests are supposed to be more efficient than their conventional counterparts. The results just reported support this supposition. A direct conventional-sequential comparison strengthened this support. The 30-item Arithmetic Reasoning (AR) component of the ASVAB provided the conventional data, and one of the corresponding Method 2 selective tests provided the sequential data for the comparison. Involving a 25% failure rate with $\alpha = \beta = .10$, the particular selective test compared had a mean length of 10 items and a 73-item frequency of 19 (see Table 2). Table 3 shows the corresponding decision-outcome percentages. The 2.5 in the lower-right cell, for example, is the overall percentage for the accepted 10% of the 25% failures ($.025 = .10 \times .25$). The selection ratio, represented by the marginal entry in the Accept column, is .70.

The base rate, complementary to the .25 failures, is .75; without testing,

Figure 2
Comparison of Observed (Open Circles) and Preset (Solid Circles) Acceptance (α) and Rejection (β) Error Rates for Three Methods of Item Selection with 25% and 75% Failure Rates, as Shown



this is the probability of selecting a potentially successful recruit. The probabilities of successful selection with testing differ markedly, not only from this value, but also from each other for the selective and conventional tests. Table 3 indicates that for the selective test the probability of successful selection is 67.5/70, or .96; the Taylor-Russell tables (Taylor & Russell, 1939) indicate by interpolation in the case of a .75 base rate and .70 selection ratio that for the AR test, with its predictive validity of .69, the corresponding probability is .88. Although the 30-item conventional test improved the probability of selecting a potentially successful recruit from .75 to .88, therefore, the improvement was notably greater for the selective test with

Table 2
Mean Test Length (Rounded) and Truncation-Item Frequency
for 1,020 Navy Recruits

Error Probability		Failure Rate			
		25%		75%	
α	β	Mean	Frequency	Mean	Frequency
Method 1					
.10	.10	12	34	10	15
.15	.10	11	26	8	7
.20	.10	10	18	6	1
.25	.10	7	9	5	1
.15	.15	8	8	5	2
.20	.15	6	5	5	0
.25	.15	4	2	4	0
.20	.20	4	1	5	0
.25	.20	3	0	2	0
.25	.25	3	0	2	0
Method 2					
.10	.10	10	19	8	8
.15	.10	8	9	6	6
.20	.10	7	8	5	1
.25	.10	7	5	4	1
.15	.15	6	4	5	1
.20	.15	5	3	4	0
.25	.15	5	2	4	0
.20	.20	5	3	4	0
.25	.20	4	1	3	0
.25	.25	4	0	3	0
Method 3					
.10	.10	20	50	15	25
.15	.10	16	30	13	11
.20	.10	13	16	11	8
.25	.10	10	10	10	4
.15	.15	14	17	11	6
.20	.15	10	7	10	4
.25	.15	8	4	9	2
.20	.20	9	1	8	1
.25	.20	7	0	5	0
.25	.25	5	0	4	0

its expected length of only 10 items: from .75 to .96. The contrast among the different selection procedures is even sharper in terms of failure, as opposed to success, probabilities. In these terms the 30-item conventional test reduced the probability of selecting a recruit who would fail from .25 to .12, while the reduction for the selective test, with its expected length of only 10 items, was from .25 to .04. Sequential testing for selection thus compares favorably on real data with conventional testing for the same purpose.

Table 3
Decision-Outcome Percentages for
Selective Test with 25% Failure
Rate and $\alpha = \beta = .10$

	Decision		
Outcome	Reject	Accept	Total
Success	7.5	67.5	75
Failure	22.5	2.5	25
Total	30	70	100

Discussion

Methods 1 and 2 appear to produce good matches of observed with expected error rates for values of α and β between .10 and .20 (see Figure 2). Discrepancies tend to appear for values larger than .20. The observed values produced for $\alpha = .25$ or $\beta = .25$ tend to approximate the observed values for $\alpha = .20$ or $\beta = .20$. One reason for this tendency may be that the mean test lengths both for $\alpha = .20$ and $\beta = .20$ and for $\alpha = .25$ and $\beta = .25$ are approximately equal, both being nearly as small as possible, so that for both pairs of expected error rates testing tends to end at about the same item number with about equal observed error rates. Another reason is that for $\alpha = \beta = .20$ and for $\alpha = \beta = .25$, for example, the corresponding critical values tend not to differ very much: .25 and 4 for $\alpha = \beta = .20$ and .33 and 3 for $\alpha = \beta = .25$. Discrepancies also tend to occur for $\alpha = .05$ or $\beta = .05$.

Though not shown in Figure 2, because the long intersecting lines would confuse the figure, the discrepancies can be quite large. For $\alpha = \beta = .05$, in the case of a 25% failure rate, for example, the corresponding Method 2 observed values were .18 and .14. Discrepancies as large as these may be due to the large truncation item frequencies typical of low expected error rates. In the case of the preceding example, with a mean test length of 16, the frequency for truncation item 73 was 68, much larger than the largest value (19) shown for Method 2 in Table 2.

The high mean test lengths and truncation item frequencies may indicate generally poor discriminability among the ASVAB items for predicting AFQT scores. In the study reported by Weitzman (1982), involving different predictor items and a different criterion, the matches for $\alpha = .05$ and $\beta = .05$ were considerably better than here.

Table 2 shows notably lower mean test lengths and truncation item frequencies for the 75% than for the 25% failure rate. This difference may be due to the breaks in the frequency distribution, shown in Figure 1, near the 75th centile (zero frequencies for AFQT scores of 72 and 74). In contrast, the frequencies all tend to be quite large around the 25th centile. Test length and error rate accuracy may thus depend on the criterion as well as on the test items used to predict it.

The local independence assumption accommodated by Method 3 appears to hold in Methods 1 and 2 without special accommodation. The Method 3 attempt to accommodate the local independence assumption, in fact, failed noticeably to reduce the discrepancies between the observed and the expected error rates. Sample size also appears to have had no noticeable effect on Method 3 matches between these error rates. The matches are no better on the right (N = 765) than on the left (N = 255) for Method 3 in Figure 2. The mean test lengths indicate further that Method 3 may not be as good as Methods 1 and 2. For $\alpha = \beta = .10$ with a 25% failure rate, for example, the Method 3 mean test length was 20. This mean test length compares unfavorably with the corresponding mean test lengths for Method 2 (10) and for Method 1 (12).

Altogether, the results for Methods 1, 2, and 3 indicate that no special attempt to meet the assumption of local independence is necessary. Simply a correlation coefficient appears to be adequate as an objective function; and in the choice between correlation coefficients, the phi coefficient (Method 2) seems preferable. Of the three objective functions studied, this coefficient, which maximizes item discriminability, generally yielded not only the best matches between observed and expected error rates but also the lowest mean test lengths and truncation item frequencies.

At least two recommendations would appear to follow from the results of this research. Prescreening applicants for military service optimally requires, for test security, that each applicant take a unique set of test items. The first recommendation is thus that item development aim at the creation of an item bank consisting of items that are more or less equally discriminating in the region of the anticipated AFQT cutting scores. Item selection from a bank like this can be random without affecting the accuracy or the length of the test. Because the AFQT is a linear combination of ASVAB tests, the AFQT frequency distribution has a number of breaks in it. The second recommendation is that the cutting score be at one of these breaks. Since the selection decision is most difficult for applicants closest to the cutting score, making the cut at a score that is not obtained by any examinee ought to facilitate selection.

REFERENCES

Armitage, P. Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. Journal of the Royal Statistical Society, 1950, 12, 137-144.

Epstein, K. I., & Knerr, C. S. Applications of sequential testing procedures to performance testing. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Ferguson, R. A model for computer-assisted criterion-referenced measurement. Education, 1970, 91, 25-31.

Kalisch, S. J. A model for computerized adaptive testing related to instructional situations. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Kingsbury, G. G., & Weiss, D. J. A comparison of ICC-based adaptive mastery testing and the Waldian probability ratio method. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Linn, R. L., Rock, D. A., & Cleary, T. A. Sequential testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.

Reckase, M. Some decision procedures for use with tailored testing. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory, 1980.

Taylor, H. C., & Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. Journal of Applied Psychology, 1939, 23, 565-578.

Wald, A. Sequential tests of statistical hypotheses. Annals of Mathematical Statistics, 1945, 16, 117-186.

Weitzman, R. A. Sequential testing for selection. Applied Psychological Measurement, 1982, 6, 337-351.

DISCUSSION

MARK D. RECKASE
AMERICAN COLLEGE TESTING PROGRAM

It is usually the role of a discussant to take a set of papers and show how they contribute to a larger coherent body of work. Unfortunately, in this case the contents of the two papers I am to discuss are so different that it is difficult to deal with them together. While they both deal with very important topics in psychometrics--the effects of multidimensionality on test results, and how to make decisions using test results--each paper is best considered separately from the other.

Weiss and Suhadolnik

The Weiss and Suhadolnik paper addresses the issue of the robustness of adaptive testing to the violation of the unidimensionality assumption required by most item response theory (IRT) models. The approach that they take is to generate simulated test data according to a multidimensional model, use that data in a unidimensional simulation of adaptive testing, and then compare the obtained θ estimates to the known θ values from the first dimension.

For studies of this type to be meaningful, the procedure used to generate the simulated item response data must be as similar to real item response data as possible. If it is not, the results of the study cannot be generalized to real testing situations. Of course, simulated test data never totally capture the richness of real test data, but we are usually willing to give up the richness in sources of variation in exchange for knowledge of the true ability of the examinees (simulees in Weiss and Suhadolnik's terminology). Whether the procedure used in this paper generates data that are "close enough" to real item response data is critical to interpreting the results of this paper. Therefore, I will concentrate most of my remarks on the data generation procedure.

Before addressing the data generation procedure directly, a definition of dimensionality is needed. To me, the dimensionality of a set of item response data is the result of a complex interaction between the characteristics of the examinees and the characteristics of the items. Examinees have many different abilities, the sum total of which define the complete latent space (Lord & Novick, 1968). These abilities may be related to each other in complex ways. The items also have many characteristics. They have different reading levels, different numbers of symbols, different lengths, and address different concepts. The response to an item is a function of the person's and item's characteristics:

$$x_{ij} = f(\underset{\sim}{\sigma}_i, \underset{\sim}{\theta}_j) \qquad [1]$$

where x_{ij} is the response to item $\underline{i}$ by person j,

$\underset{\sim}{\zeta}_i$ is the vector of item characteristics, and

$\underset{\sim}{\theta}_j$ is the vector of person characteristics.

When simulating item response data, both $\underset{\sim}{\zeta}_i$ and $\underset{\sim}{\theta}_j$ must be specified as well as the function that relates them to the response.

In this study, $\underset{\sim}{\theta}$ was generated so that the θs on each dimension were uniformly distributed between -3.2 and +3.2 and the dimensions were unrelated. In addition, for the first dimension, only 17 uniformly spaced values of θ were used, with 100 simulees at each value. This was done to allow for the computation of several indices of the quality of the adaptive testing procedure.

The initial question that comes to mind when considering this θ structure is, "Is it reasonable?" Certainly the θ distribution for actual groups of examinees is probably not uniform on any dimension, and most likely the dimensions should be somewhat related. However, for the sake of a uniform evaluation of the adaptive testing procedure, this part of the simulation is probably justified. However, the heavy "tails" of the distributions should be kept in mind when interpreting the results. The high number of high and low θs will probably result in more perfect and zero raw scores than usual.

The specification of the item characteristics for this study is much more complicated. The initial characteristics of the test items were borrowed from the factor structure of the General Science subtest of the ASVAB, presumably to ensure that they matched real items. In reviewing the structure (see their Table 1), I noticed that the loadings of the first factor tended to be inversely related to the item number. I have seen this type of pattern before when analyzing multiple-choice test items that were arranged in order of increasing difficulty. The reduction in factor loadings is due to the relative increase in guessing for the more difficult items (see Reckase, 1981, for examples using simulated test data). This is an important factor, because in order to develop the full item pool the factor loadings were reproduced six times and, I assume, they were not reordered. This means that the relationship between the factor loadings and the item difficulty was probably not maintained in the full 150-item adaptive testing item pool.

Using the 150 × 4 factor loading matrix derived by reproducing the matrix from the 25 ASVAB items six times, 45 different item pool structures were developed and item responses were generated. I will not discuss this process for all of the item pool structures, but a detailed analysis of several may prove informative.

Structure 1 is defined by the factor loadings from the first factor for the 150 simulated items. In order to generate dichotomous data to correspond to this structure, parameters were generated for each item for the 3-parameter logistic model. The $\underline{a}$ parameters were obtained from the following formula which was derived from the normal ogive model (Lord & Novick, 1968, p. 378):

$$a_{gj} = F_{gj}/(1 - F_{gj}^2)^{\frac{1}{2}} \qquad [2]$$

where a_{gj} is the a parameter estimate for item g and factor j, and F_{gj} is the factor loading for item g on factor j. The b parameters were randomly sampled from a uniform distribution from -3.2 to 3.2, and the c parameters were randomly sampled from a normal distribution with mean equal to .20 and standard deviation equal to .02.

The three parameters for each item were used to compute the probability of a correct response for the first dimension. This probability was compared to a uniform random number in the 0-1 range to determine the dichotomous response. If the random number was greater than the probability, an incorrect response was generated; if it was less than the probability, a correct response was generated.

There are two possible problems with generating the data in this way. First, the difficulty and guessing parameters do not correspond to the a parameters, and second the a parameters were based on a normal ogive model, but were used in a logistic model. Since the normal ogive as are on a different scale than the logistic as, this will produce aberrant results.

When a two-dimensional dataset was produced, the loadings on the second factor were a multiple of the factor loadings on the first factor (except in one case when the actual ASVAB factor was used). For Dataset 2, the multiplier was selected so that the second factor accounted for 1/8 the variance of the first factor. The IRT item parameters were computed in exactly the same way for the second factor as for the first factor.

To generate the multidimensional data, probabilities of a correct response were computed separately for each dimension. These were then combined using a weighted average procedure, weighting the probabilities by the squared factor loadings. This process results in a compensatory model, but the compensation is on the probability metric, not on the θ metric as it is in some other models (McKinley & Reckase, 1983).

The characteristics of the data generated by this procedure are very difficult to predict. The smaller a values for the second dimension will tend to keep the probabilities computed for that dimension close to .5, but weighting the probabilities by the squared factor loading will tend to reduce the influence of the second factor. In effect, the impact of the second factor is being reduced twice, first from the reduced a value, and second from the weighting by the factor loading.

An important point to be made about this procedure is that the magnitude of effects of the factors can no longer be described in terms of proportion of variance accounted for. Such a description is only appropriate when dimensions are combined linearly on the factor score metric. In this case, the combination of dimensions is being done on the probability metric, a nonlinear transformation of the factor score metric.

The point of describing the data generation process in detail was to demonstrate how difficult it is to determine the characteristics of the data that were produced. The process has so many complexities that it is even difficult to determine whether the resulting data are really multidimensional. It would have been informative if the authors had factor analyzed the simulated data so that the characteristics of the data could be determined. Until confidence can be gained in the data generation procedure, it is difficult to have confidence in the results of the study.

Weitzman

The Weitzman paper addresses a second important problem in the area of psychometrics: how to make decisions using test scores. In addressing this problem, he proposes a very intriguing variation of Wald's (1947) sequential probability ratio test (SPRT). Before discussing Weitzman's variation, I will first briefly describe Wald's procedure.

When it is desired to decide whether a person is above or below a particular cutting score, the SPRT requires that an indifference region be defined. This is an area on the score scale where it is a matter of indifference whether a person is classified as above or below the cutting score. The region is typically defined by its upper and lower boundaries, θ_1 and θ_0 respectively. The SPRT has error rates of α and β at the limits of the indifference region. The error rates are higher within the region and lower outside the region. The actual test statistic is the ratio of the likelihood of the observed responses $x_1, x_2, \ldots, x_n$ at the upper limit of the indifference region to the likelihood of the observed responses at the lower limit of the indifference region,

$$\frac{L(x_1,x_2,\ldots,x_n \mid \theta_1)}{L(x_1,x_2,\ldots,x_n \mid \theta_0)} . \qquad [3]$$

Weitzman proposes to do away with the indifference region by changing the form of the likelihood functions used. His test statistic can be formulated as

$$\frac{L(x_1, x_2, \ldots, x_n \mid \theta > \theta_c)}{L(x_1, x_2, \ldots, x_n \mid \theta < \theta_c)} \qquad [4]$$

where θ_c is the cutting score. His Equation 1 is a form of this statistic using quantiles to approximate the continuous functions involved.

At first glance, Expression 4 would seem to reach Weitzman's goal of eliminating the indifference region. However, the expression can be reformulated to show that it is equivalent to a classical SPRT with an indifference region.

The value of $L(x_1, x_2, \ldots, x_n|\theta > \theta_c)$ is essentially a weighted average of the $L(x_1, x_2, \ldots, x_n|\theta)$ for $\theta > \theta_c$ where the weights are the density of the distribution of θ at each θ value. Therefore, the value of the likelihood will be between $L(x_1, x_2, \ldots, x_n|\theta_c)$ and $L(x_1, x_2, \ldots, x_n|\theta \text{ max})$. Thus, for some value of θ, θ', such that $\theta_c < \theta' < \theta$ max,

$$L(x_1, x_2, \ldots, x_n \mid \theta') = L(x_1, x_2, \ldots, x_n \mid \theta > \theta_c). \qquad [5]$$

Likewise a value of θ, θ", can be found such that

$$L(x_1, x_2, \ldots, x_n \mid \theta'') = L(x_1, x_2, \ldots, x_n \mid \theta < \theta_c). \qquad [6]$$

Expression 2 can then be rewritten as

$$\frac{L(x_1, x_2, \ldots, x_n \mid \theta')}{L(x_1, x_2, \ldots, x_n \mid \theta'')} \qquad [7]$$

where θ' and θ" are the limits of the implied indifference region. It is at these two points that the specified error rates will hold. Thus, Weitzman's procedure is not really any different than Wald's SPRT. The only difference is that he does not know the extent of his indifference region, while Wald shows how to specify the region. Also, according to the procedure specified, the region will shift after each item is administered rather than remaining constant throughout the decision-making process. It would probably be better to have control over the limits of the indifference region rather than allow them to float as a function of the items selected.

Despite the problems with the procedure, Weitzman's study does show the value of SPRT types of procedures in increasing the efficiency of decision making using test data. He has also demonstrated an item selection procedure based on classical test theory that is equivalent to the IRT approach I proposed earlier (Reckase, 1983). Both of these results are a valuable contribution to the research in this area.

REFERENCES

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

McKinley, R. L. and Reckase, M. D. An application of a multidimensional extension of the two-parameter logistic latent trait model (Research Report ONR83-3). Iowa City, IA: The American College Testing Program, Resident Programs Department, 1983.

Reckase, M. D. A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press, 1983.

Reckase, M. D. The formation of homogeneous item sets when guessing is a factor in item responses (Research Report 81-5). Columbia, MO: The University of Missouri, Educational Psychology Department, 1983.

Wald, A. Sequential analysis. New York: Dover Publications, 1947.

COMPONENT LATENT TRAIT MODELS FOR TEST DESIGN

SUSAN EMBRETSON (WHITELY)
UNIVERSITY OF KANSAS

Research on aptitude tests has changed considerably in the last decade. The infusion of cognitive psychology into aptitude research has revitalized the field. Research on the cognitive components of aptitude (Carroll, 1976; Pellegrino & Glaser, 1979; R. Sternberg, 1977), as well as the cognitive correlates of aptitude (Hunt, Lunneborg, & Lewis, 1975), not only has changed the content of aptitude theory but also has influenced the type of data that is deemed relevant.

Cognitive psychology differs markedly from psychometrics on the role of the stimulus in task performance. Cognitive psychology experiments often employ within-subjects factorial designs in which stimuli are systematically manipulated to represent different levels of theoretical variables. Other theoretical variables that could influence performance are either held constant over the set or counterbalanced to eliminate bias. These experiments are like psychological tests in that many problems of a single task type are presented. However, the goal is to decompose the stimulus factors in the task that influence performance.

Cognitive component analysis of aptitude seeks to decompose the factors that influence performance on aptitude test items. A wide variety of the item types that appear on popular tests have been studied experimentally. For example, linear syllogisms (Sternberg & Weil, 1981), series completions (Butterfield, in press), and spatial problems (Pellegrino, Mumaw, & Cantony, in press), as well as many other item types, have been studied in recent research on cognitive components. The factors that have been identified on these tasks include the processes, strategies, and knowledge stores that underlie performance.

Cognitive component decomposition of aptitude offers a new approach to psychological measurement. This approach is test design, in which the qualities that are measured by a test are operationalized by the design of the test stimuli. That is, just like an experimenter who designs tasks to test hypotheses, an item writer manipulates the stimulus features of an item to represent specified theoretical constructs. Test design may be applied to many substantive areas and linked directly to psychometrics (see Embretson, in press-b).

The test design approach involves qualitatively different assumptions about the nature of construct validation research. Traditionally, the construct validity of a measure is assessed through the relationship of individual differences on the test to other measures. Recently, Embretson (in press-a) has elaborated on two separate goals in construct validation research-construct representation and nomothetic span. Embretson (in press-a) hypothesized that the shift of psychological research to structuralism permits construct representation to be studied separately from nomothetic span. In Embretson's (in press-a) conceptualization of construct validity, construct representation is assessed from task decomposition data, while nomothetic span is assessed from individual differences data. That is, the theoretical constructs that are represented in performance may be studied independently from the utility of the test as a measure of individual differences. Thus, the construct validity of the test depends, in part, on the representation of the underlying constructs in the item task.

The goal of the current paper is to present three latent trait models that can be used for test design. Estimating the parameters for these models depends on applying a method for task decomposition. Thus, prior to presenting the latent trait models, two methods for task decomposition will be presented, along with examples that illustrate their relevance for test design. Then, the three latent trait models will be presented. These are (1) the linear logistic latent trait model (Fischer, 1973); (2) the multicomponent latent trait model (Whitely, 1980d); and (3) the general component latent trait model (Whitely, 1980a). The latter is a generalization that includes the other two models. Last, the need for more complex latent trait models to fully assess the important cognitive components of aptitude will be examined. That is, the potential contribution of metacomponent latent trait models to test validity will be explored.

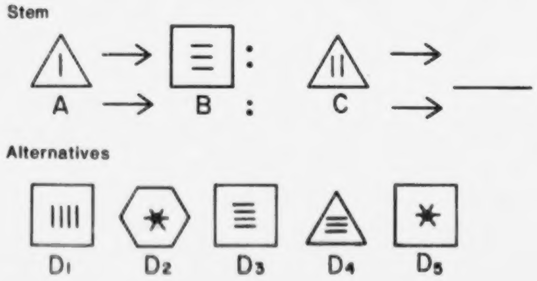
Methods for Task Decomposition

Methods for task decomposition are a major tool in contemporary research on cognitive components. The methods that are applied to decompose tasks may also be applied to the design for test stimuli. Two popular methods for task decomposition are (1) the method of complexity factors and (2) the method of subtasks. An example of how task decomposition methods can be used for test design will be presented for each method.

In the method of complexity factors, each item is manipulated and/or scored on one or more factors that represent the item's position on underlying theoretical variables. This method has been applied to attitude and personality items (Cliff, 1977; Cliff, Bradley, & Girard, 1973), as well as to a wide variety of cognitive tasks, such as linear syllogisms, geometric analogies, series completion problems, and spatial rotation items.

Figure 1 presents an example of a geometric analogy (Whitely & Schneider, 1981) that represents the method of complexity factors. Two processing events have been indicated as having major influence on task difficulty (Mulholland, Pellegrino, & Glaser, 1980; Whitely & Schneider, 1981). These are (1) encoding complexity, which depends on the number of elements in the A term in the analogy and (2) transformational complexity, which depends on the number of transformations that are required to convert A to B. In Figure 1 the A term contains two elements (the triangle and the line) and the A to B conversion requires three transformations (a shape change of the external element, an increase in the number of internal elements and a 90° rotation of the internal elements). Whitely and Schneider (1981) found that two different types of transformation had opposing influence on item difficulty. Distortions (change in shape or number) were positively related to accuracy, while displacements (rotations) were negatively related to accuracy.

Figure 1
A Geometric Analogy, Similar to an Item
on the Cognitive Abilities Test



These findings indicate that the test developer can control item difficulty by systematically varying the number of elements and the number and type of transformations in the item stimuli. An easy item would have one or two elements and a distortion transformation. A difficult item would have several elements and one or more displacement transformations. Thus, the test developer can fashion items to achieve desired levels of difficulty.

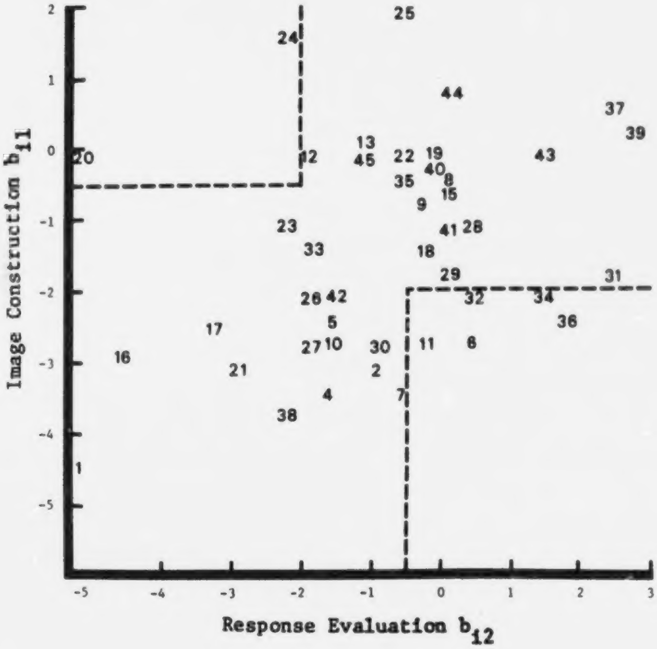
In contrast to the method of complexity factors, the method of subtask responses requires the theoretical variables to be identified from a series of subtasks that have been constructed from the items. Table 1 presents a verbal analogy item that is similar to items on the verbal section of the Cognitive Abilities Test. The total item, as presented on the test, is given at the top. Two components that have been supported by previous experimental research on verbal analogies are Rule Construction and Response Evaluation (Pellegrino & Glaser, 1979; Whitely, 1980c; Whitely & Barnes, 1979). These are represented by the two subtasks in Table 1. Notice that although Response Evaluation is sequentially dependent on Rule Construction, supplying the rule in the subtask makes possible independent assessment of these components. Thus, for each item, examinees respond to the total item as well as to the subtasks that represent processing components.

By using the psychometric models to be described below, item difficulty on the components underlying the subtasks can be calibrated on a common scale. Figure 2 presents a scatterplot of the item parameters on the two components. It can be seen that item difficulty on the two components is not highly related. Thus, it is possible to design tests that reflect predominantly the influence of one component or the other. For example, items that are easy on Response Evaluation but difficult on Image Construction would measure abilities on the latter. The test developer could select the items in the lower right corner to meet this specification.

Table 1
Subtask Set for Verbal Analogy Components

Total Item
Cat : Tiger :: Dog : __________
(a) Lion (b) Wolf (c) Bark (d) Puppy (e) Horse
Rule Construction
Cat : Tiger :: Dog : __________
Rule ?______
Response Evaluation
Cat : Tiger :: Dog : __________
(a) Lion (b) Wolf (c) Bark (d) Puppy (e) Horse
Rule: A large or wild canine

Figure 2
Scattergram of Image Construction Difficulty by Response Evaluation Difficulty for 45 Verbal Classification Items



Desiderata for Psychometrics for Test Design

The indices that are derived from classical test theory or latent trait theory do not reflect the stimulus properties of a test item with respect to specified factors. There are several desiderata for test theory models that can be applied to test design. First, the method must be capable of testing hypotheses about the specification factors. Obviously, a viable specification system is one that is highly related to item difficulty. However, hypothesis testing about the factors in items is also crucial to establishing a theory of the item task. An item specification system is implicitly a theory of the task so that it should be evaluated by the hypothesis-testing methods that are applied to other theories. Second, the model must have parameters to describe the difficulty of the items on the underlying factors. The unidimensional latent trait models that are popular in test development do not have this property, since the items are calibrated for only one dimension--the largest common factor in the items. A model that allows designation of the difficulty factors according to an a priori specification is required. Third, measurements of persons must be included in the model. The need for person measurements is self-evident, since the goal of aptitude testing is to measure individual differences. Fourth, the model should specify the relationship between the item parameters and the person abilities. Optimally, the test design approach involves selecting from a calibrated item bank for a certain measurement goal. It is essential that the influence of item parameters on person abilities is well specified in the model.

Component Latent Trait Models

This section presents three component latent trait models that can be used to test hypotheses about construct representation and to assess factors for test design. These are (1) the linear logistic latent trait model, (2) the multicomponent latent trait model, and (3) the general component latent trait model. The latter, a generalization of the other two, can handle more complex data about cognitive processes.

The Linear Logistic Latent Trait Model

The model. The linear logistic latent trait model (LLTM) is a unidimensional model in which components are identified from item scores on complexity factors that are postulated to determine item difficulty. To understand how components are identified, consider the geometric analogy presented in Figure 1, which is similar to items on the nonverbal section of the Cognitive Abilities Test. A recent study (Whitely & Schneider, 1981) compared three cognitive models of geometric analogies, using the LLTM. All three models specify complexity factors in processing the item that influence response difficulty.

The scores of the items on the complexity factors identify the components in an LLTM. The model can be examined by considering three equations. The first equation is the mathematical model for task processes. Here, a linear model of the complexity factors, c_{im}, multiplied by their difficulty, η_m, predicts item difficulty, b_i^*.

$$b_i^* = \sum_m c_{im}\eta_m + d \ , \qquad [1]$$

where

c_{im} = the complexity of factor m in item i;

η_m = the difficulty of complexity factor m; and

d = a normalization constant.

The second equation presents the latent trait model for individual differences,

BEST COPY AVAILABLE

which is the Rasch latent trait model,

$$P(x_{ij}=1|\theta_j,b_i) = \frac{e^{(\theta_j-b_i)}}{1 + e^{(\theta_j-b_i)}} \qquad [2]$$

where θ_j = ability for person j and b_i = difficulty for item i.

Equation 3 combines these two models to give the LLTM as follows:

$$P(x_{ij}=1|\theta_j,\eta_m,d) = \frac{e^{(\theta_j-(\sum_m c_{im}\eta_m + d))}}{1 + e^{(\theta_j-(\sum_m c_{im}\eta_m + d))}} \qquad [3]$$

If the number of complexity factors equals the number of items, and each item contains only one complexity factor, then the LLTM is equivalent to the Rasch latent trait model. When the number of η_m is less than the number of items, the LLTM is a linearly constrained model of item difficulty.

A major advantage of the LLTM is the possibility of comparing alternative models of item difficulty by χ^2 difference tests based on the log likelihood of the data, given the model. For example, the fit of any restricted model of the task components can be compared to the fit of the Rasch model, which can be regarded as a saturated model of item difficulty. Further, if alternative models of task components are hierarchically nested, then direct comparisons between the models are also possible. Thus, hypothesis testing to establish a valid model of the task complexity factors is an important capability of the LLTM.

Another important aspect of the model is that of parameters describing each item by component complexity rather than just item difficulty. These parameters can be useful in item banking, so that the contribution of a processing complexity factor to each item is systemically specified. Notice, however, that the model is unidimensional, since only one ability parameter is specified for each person.

Estimation. Fischer (1973) derived conditional maximum likelihood estimators for the item parameters of the LLTM, η_m. Although conditional maximum likelihood estimators are statistically superior to unconditional estimators for several reasons (Fischer, 1981), they are impractical for large sets (I > 60). Thus, unconditional maximum likelihood estimators are useful for LLTM item parameters.

The first derivative of the log likelihood function for unconditional maximum likelihood estimation is

$$\frac{\partial L}{\partial b_i} = \sum_j \left[x_{ij} - \left(1 + e^{-(\theta_j - b_i)}\right)^{-1}\right]. \qquad [4]$$

Thissen (1981) has shown that the first derivative of the log likelihood function L for the LLTM with respect to η_m may be obtained from

$$\frac{\partial L}{\partial \eta_m} = \sum_i c_{im} \left[\frac{\partial L}{\partial b_i} \right] . \qquad [5]$$

Combining Equation 4 with Equation 5 gives the first derivative of the unconditional log likelihood function with respect to η_m

$$\frac{\partial L}{\partial \eta_m} = \sum_i c_{im} \sum_j \left[x_{ij} - \left(1 + e^{-(\theta_j - b_i^*)} \right)^{-1} \right] , \qquad [6]$$

where b_i^* is defined as in Equation 1.

Multicomponent Latent Trait Model

The model. The multicomponent latent trait model (MLTM) is a multidimensional model in which components are identified from subtasks that represent the processing components in item solving. Like many information processing models for complex tasks (e.g., Hunt, 1976), it is assumed that information from several component events is required to solve the item. The relationship between the component events may be either (1) independent, where the processing or outcome of one event does not influence any other event, or (2) sequentially dependent, where information from a component event provides prerequisite information for processing on later events.

A MLTM uses subtask data to identify the components. The mathematical model of processes in the MLTM links the component responses to the total item. Equation 7 presents a mathematical model for independent components in which the response probability for the total item is the product of the component likelihood:

$$P(x_{ijT}=1) = a \prod_k P(x_{ijk}=1) + g \left[1 - \prod_k P(x_{ijk}=1) \right] . \qquad [7]$$

where

$P(x_{ijT}=1)$ = the probability that the composite task is correct for person j on item i,
$P(x_{ijk}=1)$ = the probability that the subtask for component k is correct for person j on item i,
a = the probability that an item is solved when the component information is available, and
g = the probability of solving an item when the component information is not available.

Unlike the original MLTM, the model includes parameters for application of the component information, a, which represents metacomponent or executive functioning, and for an alternative method for solving the item, g, such as guessing or rote association to the stem. Other mathematical models are possible (e.g., Whitely, 1980d), but all models relate the component likelihoods to the full item likelihood.

As for the LLTM, the latent trait model for individual differences is the 1-parameter logistic latent trait model, as presented in Equation 2. However, in the MLTM the latent trait models are given for component subtask responses rather than for the total item. The MLTM specifies that responses to the subtasks depend on the ability of person j on component k and the difficulty of item i on component k, as follows:

$$P(x_{ijk}=1|\theta_{jk},b_{ik}) = \frac{e^{(\theta_{jk}-b_{ik})}}{1+e^{(\theta_{jk}-b_{ik})}}\ , \qquad [8]$$

where θ_{jk} = the ability of person j on component k and b_{ik} = the difficulty of item i on component k. The LLTM, in contrast, is a latent trait model for responses to the total item and does not model component responses.

The full model, presented in Equation 9, combines the latent trait model with the mathematical model. It can be seen that the total item response is conditional on K component abilities as well as on K component item difficulties.

$$P(x_{ijT}=1|\underset{\sim}{\theta}_j,b_i) = (a-g)\prod_k\frac{e^{(\theta_{jk}-b_{ik})}}{1+e^{(\theta_{jk}-b_{ik})}}+g\ . \qquad [9]$$

where $\underset{\sim}{\theta}_j$ = the vector of k component abilities for person j and b_i = the vector of k and component difficulties for item i.

Although typical test data (with the notable exception of linked items with a common stem) only occasionally assess subtask responses, there are several reasons why it may be useful to obtain such data as part of test development. First, the various processing components from which information is required for item solution are theoretically distinct. Experimental cognitive research has supported the independence of components within a task by additive factor (S. Sternberg, 1969) or subtractive factor (Pachella, 1974) modeling methods. Second, if the components are sufficiently elementary, they should generalize across tasks and possibly account for differential patterns of correlations in performance on separate types of items (Carroll, 1974). Third, individual differences on different components correlate only moderately and show differential validity in predicting performance on other tasks (R. Sternberg, 1977; Whitely, 1981). Fourth, component difficulties are sometimes not highly correlated in item sets, so that it is possible to select items of the same type that measure different component abilities. Consider for example, a two-component item, such as presented in Table 1. If items are so easy on one component that nearly everyone has a high probability of executing it correctly, then it can be shown that the likelihood of correctly answering the items is well described by the regression of the response likelihoods on the other component ability (Whitely, 1981).

The relationship of the MLTM parameters to the joint response to the components, C_i, and the total item T, is explicated more completely by considering the probability sample space. There are 2^{K+1} possible response patterns. Table 2 shows the eight response patterns for a two-component item, along with an expression for the probability of the pattern from the MLTM. It can be seen that the a and g parameters link the component response to the total item, while the other symbols represent the probability of the component response patterns, which vary systematically over persons and items, according to the 1-parameter logistic latent trait model.

Estimation. No estimators were developed in Whitely (1980d) for the MLTM. However, given a probability space such as specified in Table 2, the likelihood of any response pattern is given by

$$P(x_k,x_T)=\left[g^{x_T}(1-g)^{1-x_T}\right]^{1-\Pi_k x_k}\left[a^{x_T}(1-a)^{1-x_T}\right]^{\Pi_k x_k}\left[\prod_k P_{x_k}^{x_k} Q_{x_k}^{1-x_k}\right] , \qquad [10]$$

where

x_k = vector of responses of person j to components for item i,

x_T = response of person j to total item i, and

P_{x_k} = $P\,(x_{ijk}=1\,|\,\theta_{jk},\ b_{ik})$.

Notice that the entry of the parameters a and g into the likelihood depends on the value of $\pi_k x_k$ and that $\pi_k x_k$ equals 1.0 only if all component outcomes are correct. Note also that a contributes to the log likelihood only if all the components are executed correctly, while g contributes when at least one component is incorrect. This pattern is specified in Table 2.

The likelihood of the data set can be obtained by multiplying the response likelihoods over persons and items. Since neither a nor g vary over persons or items, it can be concluded immediately from well-known theorems on the binomial distribution that their maximum likelihood estimators are the relative frequencies

$$a = \frac{\sum_j\sum_i (\Pi_k x_k)\, x_T}{\sum_j\sum_i \Pi_k x_k} \qquad [11]$$

and

$$g = \frac{\sum_j\sum_i (1 - \Pi_k x_k)\, x_T}{\sum_j\sum_i (1 - \Pi_k x_k)} \ . \qquad [12]$$

Thus, a is given by the relative frequency of correctly answering the item when all components are executed correctly, while g is given by the relative frequency of correctly answering the total item when at least one component is executed incorrectly.

Table 2
Frequencies and Conditional Probabilities for
Joint Response Patterns on Verbal Analogies

C_1	C_2	T	f	$P(x_T = 1 \mid x_k)$	Notation
1	1	1	1864	.84	$a\ P_{x_1} P_{x_2}$
1	1	0	351	.16	$(1-a)\ P_{x_1} P_{x_2}$
0	1	1	518	.50	$g\ Q_{x_1} P_{x_2}$
0	1	0	518	.50	$(1-g)\ Q_{x_1} P_{x_2}$
1	0	1	84	.45	$g\ P_{x_1} Q_{x_2}$
1	0	0	101	.45	$(1-g)\ P_{x_1} Q_{x_2}$
0	0	1	87	.28	$g\ Q_{x_1} Q_{x_2}$
0	0	0	221	.72	$(1-g)\ Q_{x_1} Q_{x_2}$

$$P_{x_k} \equiv P(x_{ijk}=1 \mid \theta_{jk}, b_{ik})$$

$$P_{x_k} = \frac{e^{(\theta_{jk}-b_{ik})}}{1 + e^{(\theta_{jk}-b_{ik})}}$$

$$P(x_T=1) = a\prod_k P_{x_k} + g\left[1-\prod_k P_{x_k}\right]$$

$$P(x_k, x_T) = \left[g^{x_T}(1-g)^{1-x_T}\right]^{1-\prod_k x_k} \left[a^{x_T}(1-a)^{1-x_T}\right]^{\prod_k x_k} \left[\prod_k P_{x_k}^{x_k} Q_{x_k}^{1-x_k}\right]$$

The required derivative of the log likelihood for unconditional maximum likelihood estimation of the item parameters b is

$$\frac{\partial L}{\partial b_{ik}} = \sum_{j}\left[x_{ijk} - \left(1 + e^{-(\theta_{jk} - b_{ik})}\right)^{-1}\right] . \qquad [13]$$

Setting the derivative to zero leads to the well-known equations for unconditional maximum likelihood estimation of the 1-parameter logistic latent trait model (cf. Lord & Novick, 1968). As for other exponential families of distributions, estimation equations for the latent trait model can be obtained by equating the observed sufficient statistics with their expectancies, given the parameters (Andersen, 1980). In the current development, however, estimation requires I equations for each of K components of the MLTM. Notice that the item parameters for each component b_{ik} involve only the responses to the relevant subtask data, x_{ijk}. It can be seen that unconditional maximum likelihood estimators may be obtained independently from each subtask to maximize the log likelihood of the joint response pattern x_k, x_T.

A General Multifactor Latent Trait Model

The model. The preceding developments have shown that the LLTM and the MLTM differ substantially in component identification. LLTMs estimate difficulty of complexity factors that are related to item difficulty, while MLTMs estimate item and person parameters for component outcomes. The different methods of component identification make possible a meaningful unification of these two models.

Consider the verbal analogy that is presented in Table 1. This analogy was presented previously with the MLTM. Although the Response Evaluation component is identical to the previous example, the Rule Construction component is postulated to be influenced by the several processing complexity factors, c_{ikm}, that are listed in Table 3. These factors concern the difficulty of inferring the target relationship (i.e., Fist: Clench). The factors c_{i11} and c_{i12} are the ease of inferring the target relationship in the initial encoding of the relational pair and in the context of the unmatched term "Teeth," respectively. Previous research on analogies (R. Sternberg, 1977) as well as research in memory organization (Reitman, 1965) suggest that relational span is also positively related to item solving, since extraneous relationships can interfere with solving the analogy. In the current example, the factors c_{i11} and c_{i12} are measured by the mean number of relationships that are educed between the word pair when presented alone and in the context of the unmatched term, respectively. The factors c_{i14} to c_{i17} represent the relative frequency of various types of context effects in inferring the target relationship (i.e., selecting or combining initial relationships, inferring new relationships, and so forth).

Scores for each item on the complexity factors were obtained from other research studies on analogies (Embretson & Curtright, 1981). However, it is

Table 3
Complexity Factors in Analogical Reasoning Components

Rule Construction Component
Fist:Clench::Teeth: ____________________
Rule? ____________________

c_{ikm} = the complexity of factor $\underline{m}$ for component $\underline{k}$ on item $\underline{i}$
c_{i11} = inference elicitation, the probability that the target relationship is educed from initial word pair
c_{i12} = relational network span, the number of relationships educed
c_{i13} = inference contextualization, the probability that the target relationship is educed in context of all three stem stimuli
$c_{i1T} - c_{i17}$ = type of contextualization effect

Response Evaluation Component
Fist:Clench::Teeth: ____________________
(1) Pull (2) Brush (3) Grit (4) Gnaw (5) Jaw
Rule: Angry reaction done with "teeth."

---

important to note that in this example the complexity factors have effects on the *component* information outcomes rather than on the total item response.

Equation 14 presents a model that specifies both processing complexity factors and processing component outcomes.

$$P(x_{ijT}=1|\underset{\sim}{\theta}_j,\underset{\sim}{\eta}_m,\underset{\sim}{d}) = \prod_k \left[\frac{e^{(\theta_{jk}-(\sum_m c_{imk}\eta_{mk} + d_k))}}{1 + e^{(\theta_{jk}-(\sum_m c_{imk}\eta_{mk} + d_k))}} \right] \qquad [14]$$

As for the MLTM the probability of the correct response to the intact item is conditional on a vector of component abilities, $\underset{\sim}{\theta}_i$, and component item difficulties. However, item difficulty for each component η_m is determined by a linear model of the complexity factors for the component c_{ikm}.

Equation 14 is a general multifactor latent trait model (GLTM) for response processes. If only one information outcome is measured for the item (i.e., the response to the intact test item), then the model is identical to the LLTM. In this case, complexity factors would be scored for the total item. The parameters $\underline{a}$ and $\underline{g}$ drop out of the model, since the response to the total item would be given by response to the single component outcome that is observed. If no complexity factors postulated for each component outcome, but several component outcomes are observed, then the model is MLTM. In this case, the Rasch model is specified for each component. However, for tasks with multiple information outcomes and processing complexity factors that influence these outcomes, the full model can be utilized.

Estimation. The likelihood of the joint response pattern, x_k, x_T, given the parameters of the model, is given by Equation 10, except that the component likelihoods are given by the LLTM for the component as follows:

$$P_{x_k} \equiv P(x_{ijk}=1 \mid \theta_{jk}, \eta_{mk}, d_k) \ . \qquad [15]$$

The estimators of $\underline{a}$ and $\underline{g}$ for the GLTM are the same as for MLTM, as given in Equation 11 and Equation 12. Since item difficulty is linearly constrained within a component for the LLTM, the first derivative of the log likelihood function with respect to η_{mk} is required for unconditonal maximum likelihood estimation of the items. Using the development given above for LLTM, it can be seen that the first symbolic partial derivative with respect to η_{mk} in GLTM is

$$\frac{\partial L}{\partial \eta_{mk}} = \sum_i c_{imk} \sum_j \left[x_{ijk} - \left(1 + e^{-(\theta_j - b_i^*)} \right)^{-1} \right] . \qquad [16]$$

A Fortran program, MULTICOMP (Whitely & Nieh, 1981) is available to estimate the parameters of the GLTM.

Future Directions: Metacomponent Latent Trait Models?

The component models that were presented above do not fully reflect the complexity of the information processes that are involved in task performance. Metacomponent variables that determine when to execute component processes and which processes to execute have great impact on problem solving. For example, problem-solving strategies are an important concept in problem-solving theory (Davis, 1973; Newell & Simon, 1978). Similarly, problem-solving strategies have long been thought to be major aspects of individual differences in intelligence, particularly for those theories of intelligence that emphasize adaptability (Pintner, 1921; R. Sternberg, 1979; Woodrow, 1921). Thus, on theoretical grounds, a complete model of information processing on intelligence test items should include strategy variables.

The MLTM that is presented in Table 2 postulates that individuals have equal likelihoods of applying the various strategies. That is, the strategy application parameter $\underline{a}$, and the parameter for successful application of other strategies, $\underline{g}$, do not vary over persons or items. As suggested above, this assumption is unwarranted on both theoretical and empirical grounds, since metacomponents are known to influence task performance. Thus, to fully represent processing, the strategy application parameters need to vary over persons or items.

A metacomponent latent trait model would include strategy application parameters for persons or items. However, estimation of these parameters will be complex. Returning to Table 2, it should be obvious that the symbolic partial derivative with respect to $\underline{a}$, for example, will not be simple if variability for either persons or items is included in the model. That is, the estimation of such parameters will depend on the outcome of the other parameters, θ_j and b_i.

The parameter a can only be estimated from response patterns in which both components are correct, as in the first two response patterns in Table 2. Not only will the estimation algorithm necessarily be complex, but also estimation error will vary as a function of the other component responses. For example, for persons with few accurate component outcomes, the a parameter will not be estimated reliably, since little information about the parameter will be available.

An obvious question at this point is the potential utility of developing the estimators for the more complex metacomponent models. Two questions about metacomponent parameters need to be addressed: (1) Do items and persons vary in propensity for applying the various components? and (2) Do individual differences in metacomponents contribute to the criterion-related validity of an aptitude test? To answer these questions, data from two studies on verbal aptitude (Whitely, 1980, 1982) were reanalyzed to include the metacomponent parameters.

A reanalysis of data originally collected by Whitely (1980) shows the variability among items and persons in two metacomponents that could be estimated with the latent trait model in Table 2. These are application of the rule-oriented strategy, a, and application of other strategies, such as guessing, g. Figure 3 shows frequency distributions of the a parameters for two item types, verbal analogies, and verbal classifications. In this analysis, a was computed as the conditional probability that examinees would solve the total item when the component information was available. Admittedly, this estimator of a is crude, but it does provide at least some indication of its nature. It can be seen in Figure 3a that the parameter values tend to be high on both item types but that examinees vary widely on the parameter. It is not clear to what extent this distribution reflects differing degrees of accuracy of estimating a for individuals.

Figure 3b shows the distribution of the g parameter for examinees, computed as the conditional probability of solving the total item, given that the component information is not available. It can be seen that this value centers around .50 for both item types and that individuals vary widely in these values. Thus, for both strategy application and guessing, some individual differences are indicated. Figure 4 and Figure 5 present stem and leaf distributions of a and g parameters, respectively, for items. As for individuals, considerable variability is indicated.

A second study (Whitely, 1982) contains data on the contribution of metacomponent variables to test validity. The Whitely study examines the relationship of individual differences in strategy application to a major criterion for aptitude test validity, educational achievement. In this study, data on the achievement of 99 parochial high school students were collected, in addition to their performance on an analogical reasoning test and on several subtasks that represented components and metacomponents in solving analogies.

The contribution of strategy application parameters (i.e., a) and other strategies (g) were examined in separate analyses. In the Whitely (1982) analyses, individual differences in strategy application were examined for two strategies that led to analogy solving. There were (1) a rule-oriented strategy and (2) a response elimination strategy. The contribution of the strategy applica-

Figure 3
Frequency Distribution of Application (a) and Guessing (g) Probabilities for Examinees on Two Item Types

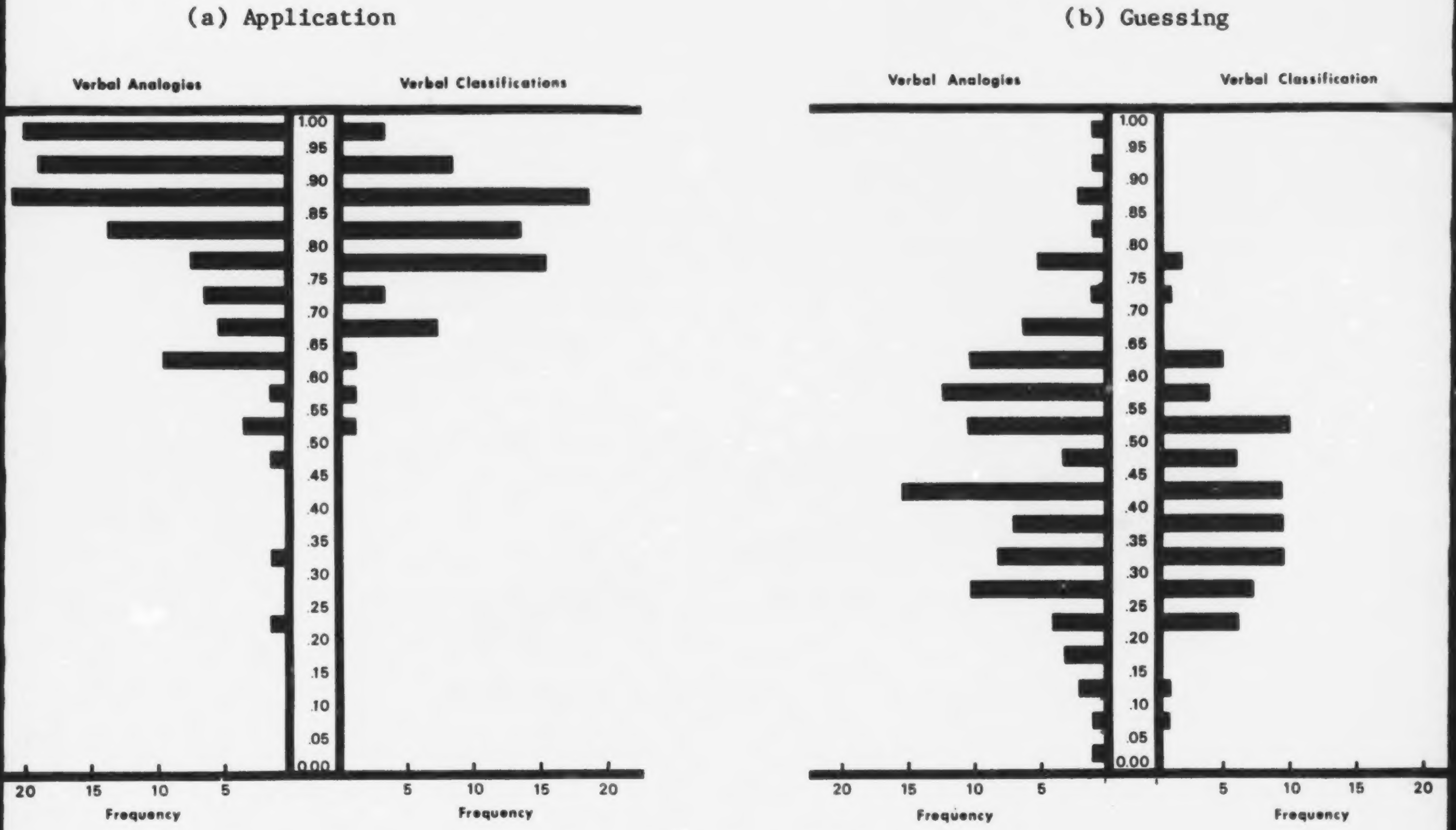


Figure 4
Stem and Leaf Distribution of Application (a)
Probabilities for Two Item Types

Verbal Analogies		Verbal Classifications
	1.000	006 028 039
030 029	.950	003 030
028 024 021 019 011 010 001	.900	002 011 016 021 023 042 043
033 032 025 022 018 017 014 009 002	.850	005 017 018 024 027 038
026 020 015 013 007 005 004	.800	004 009 013 015 019 026
034 016 003	.750	029 031 034 036 040
036	.700	008 035
027 023	.650	012 032 044
	.600	014 033 045
006	.550	020 022
037 031 012	.500	025
	.450	010
	.400	041
	.350	
	.300	
	.250	037
	.200	
	.150	
	.100	
	.050	
	0.000	

tion parameter to test validity was examined by structural equation models (Jöreskog, 1974).

In these models, individual differences in both applying and performing the components of the strategies were measured as independent variables. The dependent variables included performance on the analogical reasoning test as well as scores on eight area achievement tests. For both the rule-oriented strategy and the response elimination strategy, it was found that adding strategy application to the strategy performance variables significantly increased prediction of both analogical reasoning and achievement. The differences that were obtained by adding the strategy application variables to the covariance models were highly significant for both the rule-oriented ($X^2 = 65.57$, $p < .01$) and the response elimination strategy ($X^2 = 65.57$, $p < .01$). For both strategies the application

Figure 5
Stem and Leaf Distribution of Guessing (g)
Probabilities for Two Item Types

Verbal Analogies		Verbal Classifications
	1.000	
	.950	
	.900	
	.850	021
030	.800	003 006 011
025 021 011	.750	004
029 028 022 017	.700	002 042
018 009 002	.650	008 012 013 030 031
034 010 007 003	.600	028 036
001	.550	005
024	.500	017 029 034 038
015	.450	
033 012	.400	010 023 026
036 020 019 016 014 005	.350	009 018 019 035 043
031 023 013 004	.300	015 022 039
	.250	014 027 040
	.200	016
032 026	.150	045
006	.100	037
037 027	.050	020 024 044
	0.000	025 032 033 041

variable was significantly related to analogical reasoning ($t = 7.58$ and 2.07, respectively, for the rule-oriented and response elimination strategies), showing that strategy application is an important metacomponent for individual differences in analogical reasoning.

Table 4 presents data on the contribution of the application parameters to the prediction of achievement in several areas. Indices that are comparable to multiple regression analyses were obtained from the structural equation analyses. For each of the strategies and for the two strategies combined with a guessing strategy, Table 4 shows the F value for the application metacomponent and its incremental contribution to explaining variance of each achievement test, as well as the proportion of variance explained. The application metacomponent for the rule-oriented strategy significantly contributed to the validity for predicting Mathematics and Sources. The application metacomponent for the response elimination strategy significantly increased prediction for several achievement areas, including Reading Comprehension, Vocabulary, Language Use, Spelling, Social Science, Science, and Sources.

Table 4
Contribution of Metacomponent and Strategy Parameters to Predicting Achievement

Strategy and Achievement Area	Specification Accuracy Multiple R	Contribution of Metacomponent	
		F_β	Reduction of Error (ΔR^2)
Rule-Oriented Strategy			
Reading Comprehension	.67	.28	.01
Vocabulary	.50	.02	.01
Language Use	.67	2.01	.02
Spelling	.38	.05	.00
Mathematics	.52	4.84*	.06
Social Science	.51	1.48	.02
Science	.74	.01	.00
Source	.66	9.18**	.09
Response Elimination Strategy			
Reading Comprehension	.71	5.42*	.05
Vocabulary	.66	14.21**	.14
Language Use	.73	6.81*	.06
Spelling	.52	8.35**	.11
Mathematics	.50	.55	.01
Social Science	.57	8.58**	.10
Science	.77	5.71*	.04
Source	.71	17.64**	.15
Guessing Strategy			
Reading Comprehension	.60	9.93**	.12
Vocabulary	.63	8.44**	.09
Language Use	.58	5.50*	.07
Spelling	.48	.05	.00
Mathematics	.69	2.61	.03
Social Science	.60	6.80*	.08
Science	.71	15.20**	.14
Source	.68	10.02**	.10

*p <.05.
**p <.01.

Table 4 also shows that adding the g parameter to the rule-oriented strategy and the response elimination strategy significantly increased the prediction of achievement in several areas. Thus, these data support the potential of individual differences in metacomponent variables as an important aspect of test validity. The metacomponent variables increased the prediction of achievement over the simple component performance variables.

The main conclusion to be drawn from these studies is that metacomponent latent trait models are needed to estimate more fully the processing abilities that underlie aptitude. Although the estimation of the metacomponent parameters will be complex, even the crude estimators that were used in the studies described above show clear contributions to aptitude test validity.

Conclusions

This paper has presented latent trait models that can be used for test design in the context of a theory about the variables that underlie task performance. Examples of methods for decomposing and testing hypotheses about the theoretical variables in task performance were given. The methods can be used to determine the processing components that are involved in item performance.

Three component latent trait models for underlying theoretical variables were described along with their maximum likelihood estimators. The item parameters can be used for item banking, according to the influence of the underlying processing variables on item difficulty. Such estimators permit the test developer to choose items that represent specified information processing demands for the examinee. That is, the test developer can select items that are difficult on some processes, but easy on others. In this manner, what is measured by an aptitude test can be explicitly designed by specifying difficulty levels in the underlying processing components.

The need for metacomponent latent trait models was also considered. It was shown that both items and persons vary on metacomponent parameters and that these parameters are important for the predictive validity of an aptitude test. Thus, metacomponent latent trait models should provide a better estimate of the abilities that are involved in test performance.

REFERENCES

Andersen, E. B. *Discrete statistical models with social science applications*. Amsterdam: North-Holland Publishing Company, 1980.

Barnes, G. M., & Whitely, S. E. Problem structuring processes for ill-structured verbal analogies. *Memory and Cognition*, 1981, *4*, 411-421.

Butterfield, E. A model that predicts series completion accuracy. In S. Embretson (Whitely), (Ed.), *Test Design: Contributions from psychology, education, and psychometrics*. New York: Academic Press, in press.

Carroll, J. B. Psychometric tests as cognitive tasks: A new "structure of intellect." In L.B. Resnick (Ed.), *The nature of intelligence*. Hillsdale NJ: Erlbaum, 1976.

Cliff, N. Further study of cognitive processing models for inventory response. *Applied Psychological Measurement*, 1977, *1*, 41-49.

Cliff, N., Bradley, P., & Girard, R. The investigation of cognitive models for inventory response. Multivariate Behavioral Research, 1973, 8, 407-425.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Davis, G. A. Psychology of problem solving: Theory and practice. New York: Basic Books, 1973.

Embretson, S. E. Test design: Contributions from psychology, education, and psychometrics. New York: Academic Press, in press. (b)

Embretson, S. E., & Curtright, C. Problem structure and response format on verbal analogies (Technical Report No. NIE 81-5). Lawrence: University of Kansas, Department of Psychology, December 1981.

Fischer, G. H. The linear logistic model as an instrument in educational research. Acta Psychologia, 1973, 37, 359-374.

Fischer, G. H. On the existence and uniqueness of maximum likelihood estimates in the Rasch model. Psychometrika, 1981, 46, 59-77.

Hunt, E. Mechanics of verbal ability. Psychological Review, 1978, 85, 109-130.

Hunt, E., Lunneborg, C., & Lewis, J. What does it mean to be high verbal? Cognitive Psychology, 1975, 7, 194-227.

J$reskog, K. G. Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology. San Francisco: Freeman, 1974.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. Components of geometric analogy solution. Cognitive Psychology, 1980, 12, 252-284.

Newell, A., & Simon, H. A. Human problem solving. Englewood Cliffs NJ: Prentice-Hall, 1972.

Pellegrino, J. W., & Glaser, R. Cognitive correlates and components in the analysis of individual differences. Intelligence, 1979, 3, 187-214.

Pellegrino, J. W., Mumaw, R. J., & Cantoni, V. J. Analyses of spatial aptitude and expertise. In S. Embretson (Ed.), Test design contribution from psychology, education, and psychometrics. New York: Academic Press, in press.

Pintner, R. Intelligence and its measurement: A symposium. Journal of Educational Psychology, 1921, 12, 123-147; 195-216.

Pachella, R. G. The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), *Human information-processing: Tutorials in performance and cognition*. Hillsdale NJ: Erlbaum, 1974.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.

Reitman, W. *Cognition and thought*. New York: Wiley, 1965.

Sternberg, R. J. *Intelligence, information processing, and analogical reasoning*. Hillsdale NJ: Erlbaum, 1977.

Sternberg, R. J. The nature of mental abilities. *American Psychologist*, 1979, *34*, 214-230. (b)

Sternberg, R. J., & Weil, E. M. An aptitude strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology*, 1981, *73*.

Sternberg, S. Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, 1969, *4*, 421-457.

Whitely, S. E. *A general multicomponent latent trait model for aptitude processes* . Paper presented at the annual meeting of Psychometric Society, Chapel Hill NC, May 1980. (a)

Whitely, S. E. *Alternative strategy models for analogical reasoning*. Paper presented at the annual meeting of the Psychometric Society, St. Louis MO, November 1980. (b)

Whitely, S. E. Latent trait models in the study of intelligence. *Intelligence*, 1980, *4*, 97-132. (c)

Whitely, S. E. Multicomponent latent trait models for ability tests. *Psychometrika*, 1980, *45*, 479-494. (d)

Whitely, S. E. Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement*, 1981, *18*, 67-84.

Whitely, S. E. *The construct representation and nomothetic span of verbal ability*. Paper presented at the annual meeting of the American Psychological Association, Washington DC, August 1982.

Whitely, S. E., & Barnes, G. M. The implications of processing event sequences for theories of analogical reasoning. *Memory and Cognition*, 1979, *7*, 323-331.

Whitely, S. E., & Nieh, K. *Program MULTICOMP*. Unpublished manuscript, University of Kansas, Lawrence KS, 1981.

Whitely, S. E., & Schneider, L. M. Information structure on geometric analogies: A test theory approach. Applied Psychological Measurement, 1981, 5, 383-397.

Woodrow, H., Intelligence and its measurement: A symposium. Journal of Educational Psychology, 1921, 12, 123-147; 195-216.

ACKNOWLEDGMENTS

The author would like to thank Lisa Schneider for preparing the figures and for her help in analyzing the data. This research was partially supported by National Institute of Education Grant No. NIE 6-7-1056 to Susan E. Whitely, Principal Investigator. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Discussion: Multidimensional Latent Traits and Plausible Assumptions of Cognitive Psychology

Douglas H. Jones
Advanced Statistical Technologies Corporation

Embretson has described an application of item response theory to the current state of cognitive psychology. She explores two main cognitive processes for item response called *complexity factors* and *component tasks*. The data observed for the cognitive analysis are the item response, the level of each complexity factor for each item, and the component response for each component task on each item.

Embretson proposes a model for each set of response data depending on which cognitive process--complexity factors or component tasks--is under investigation. A third model is proposed, nesting a complexity factors model within each component task. The primitive model in all three models is the Rasch item response curve, which (as Embretson shows) facilitates the estimation of the latent parameters.

I will explore three mathematical assumptions that Embretson uses. The first concerns the complexity factors cognitive process, and the other two concern the component tasks cognitive process.

Assumption 1: The trait is unidimensional for the complexity factors model.

Suppose we consider a geometric analogy item having two complexity factors--number of elements and orientation of the object. In addition, suppose that the level of each factor is denoted c_1 and c_2. It would seem that cognitive psychologists would be interested in estimating the individual differences in respondents' abilities to mentally manipulate several objects while altering their orientation. Suppose we denote these two abilities by θ_1 and θ_2, respectively. Embretson does not account for these abilities, although it would be easy to extend her model to the following form, which we will call a quadratic-compensatory model:

$$\begin{aligned}\log \frac{P_i(\theta)}{1-P_i(\theta)} &= \eta_0 + \eta_1(\theta_1-c_{i1}) + \eta_2(\theta_2-c_{i2}) \\ &\quad + \eta_3(\theta_1-c_{i1})(\theta_2-c_{i2}) \\ &\quad + \eta_4(\theta_1-c_{i1})^2 + \eta_5(\theta_2-c_{i2})^2 \quad . \qquad [1]\end{aligned}$$

Note that this model accounts for the interaction between the levels of the two traits as well as the interaction between difficulty levels of the complexity factors. These interactions can describe the compensatory action of abilities. That is, a respondent with low ability for manipulating large numbers of objects but with high ability for rotating them could score as well as another respondent with mediocre abilities for both cognitive components.

Estimation of traits and item parameters would be a simple extension of the methods in the unidimensional case. Log-likelihood ratio statistics would be used for testing hierarchies of models.

Assumption 2: The component tasks model assumes that the conditional distribution of the total item response, given the values of the component outcomes, is independent of ability and difficulty parameters.

Factorization of the Likelihood

The primary reason for this assumption is to separately estimate the parameters (a,g) and $(\underset{\sim}{\theta},\underset{\sim}{b}_k)$. (We are using the same notation as Embretson.) The a parameter is the probability of a correct response based on the information in the component rules, and the g parameter is the probability of guessing a correct response. The components of the θ vector are the abilities for the components tasks, and the components of the $\underset{\sim}{b}_k$ vector are the difficulties of the component tasks. Letting x_T and $\underset{\sim}{x}_k$ denote the responses to the total item and component tasks, respectively, the likelihood concerning one item can be factored as follows:

$$P(x_T,\underset{\sim}{x}_k|a,g,\underset{\sim}{\theta},\underset{\sim}{b}_k) = P(x_T|\underset{\sim}{x}_k,a,g,\underset{\sim}{\theta},\underset{\sim}{b}_k)\ P(\underset{\sim}{x}_k|a,g,\underset{\sim}{\theta},\underset{\sim}{b}_k)$$

$$= P(x_T|\Pi x_k,a,g)\ P(\underset{\sim}{x}_k|\underset{\sim}{\theta},\underset{\sim}{b}_k)\ . \qquad [2]$$

The last line follows from Assumption 2, and the minor assumption that the conditional distribution of x_T depends only on Πx_k, the product of the component outcomes.

Equation 2 means that the parameters (a,g) and $(\underset{\sim}{\theta},\underset{\sim}{b}_k)$ are estimated separately. Equation 2 also means that the total item response, x_T, does not provide any useful information for estimating component task abilities as long as Πx_k is given. This is a little uncomfortable, since we feel that x_T does, on the contrary, provide information about some kind of ability no matter what the respondent does on the component tasks. Thus, the desire to separate the estimation chores leads to the need to invent more traits to explain the total item responses other than the component task traits. This is to say that the parameters $(\Pi x_k,a,g)$ are sufficient to explain the dependencies between total item responses, which means that the total item responses would not satisfy local independence.

Embretson notes this difficulty as well, offering the notation of strategy application parameters. The estimation of these parameters, contrary to Embretson's observation, will not depend on the outcome of the other parameters, $\underset{\sim}{\theta}$ and $\underset{\sim}{b}_k$, as long as Equation 2 holds. Consequently, the estimation problems would not be as complex as thought, provided strategy application parameters are introduced into $P(x_T | \Pi x_k, a, g)$ simply enough.

Item Characteristic Function

A secondary reason for Assumption 2 is for a simple derivation of the marginal distribution of x_T given $\underline{a}$, $\underline{g}$, $\underset{\sim}{\theta}$, and $\underset{\sim}{b}_k$, the item characteristic function (ICC). In general, the marginal distribution of x_T can be written as

$$P(x_T=1 | a,g,\underset{\sim}{\theta},\underset{\sim}{b}_k) = \left\{ P(x_T=1 | \Pi x_k=1) - P(x_T=1 | \Pi x_k=0) \right\} P(\Pi x_k=1) + P(x_T=1 | \Pi x_k=0) \qquad [3]$$

where parameter notation on the right-hand side has been suppressed in order to display the importance of the role played by Πx_k. Under Assumption 2, $P(\Pi x_k = 1)$ depends solely on $\underset{\sim}{\theta}$ and $\underset{\sim}{b}_k$, and the other items depend solely on $\underline{a}$ and $\underline{g}$. Thus, the ICC of x_T depends quite simply on the ICC of Πx_k.

The ICC of Πx_k will be denoted by $P(\underset{\sim}{\theta},\underset{\sim}{b}_k)$ and Equation 3 rewritten as

$$P(x_T=1 | a,g,\underset{\sim}{\theta},\underset{\sim}{b}_k) = (a-g)\, P(\underset{\sim}{\theta},\underset{\sim}{b}) + g \quad . \qquad [4]$$

This formula is interesting, since it is related to previously formulated ICCs for item responses. However, the formula is almost independently arrived at by considering the cognitive processes underlying item responses. How it is related to previous ICC formulations is as follows. For instance, for $a = 1$, $\underset{\sim}{\theta} = \theta$ (a unidimensional trait), and $P(\theta,b)$ (a 2-parameter logistic response function), Equation 4 becomes the popular 3-parameter logistic model. Moreover, recently interest has been expressed in a 4-parameter logistic model, such as Equation 4, to account for incorrect responses to easy items from high ability respondents (Barton & Lord, 1981).

Assumption 3: Local independence over component task outcomes.

The major problem with this assumption is that it demands that the total item ICC in Equation 4 be a multiplicative function of the components of θ. The model is therefore multiplicative, since the explicit equation given in Embretson can be written as

$$P(x_T=1 | a,g,\underset{\sim}{\theta},\underset{\sim}{b}_k) = (a-g) \prod_k P_{ik}(\theta_k) + g \quad , \qquad [5]$$

where

$$P_{ik}(\theta) = \frac{e^{(\theta_k - b_{ik})}}{1 + e^{(\theta_k - b_{ik})}} \quad . \qquad [6]$$

There are two reasons that Assumption 3 can cause a problem. The first reason is that the multiplicative term in Equation 5 runs counter to current research about models employing multidimensional traits. The second reason is that the assumption of local independence over task outcomes is not unreasonable. This is to say that invoking the axiom of local independence to describe the way a latent trait explains the manifest relation between task outcomes rests on the foundation of current theory (Lazarsfeld & Henry, 1968).

Thus, Embretson's paper brings forth an idea that is applicable to mainstream item response theory where no outcome task data is actually observed. When researchers try to formulate ICC models for multidimensional latent traits, they could (and should) try to imagine that each component of the latent trait relates to a task that may be performed by the respondent when he or she attempts to respond to the total item. Then the researcher would be led to adopt the local independence assumption for the outcomes of these imagined (or latent) tasks.

Conclusion

This discussion explicitly states the three main assumptions made in the Embretson paper and attempts to explore the primary results that these assumptions imply. The three assumptions are (1) the trait is unidimensional in the complexity factors model; (2) the conditional distribution of the total item response, given the values of the component outcomes, is independent of ability and difficulty parameters; and (3) the component task outcomes are locally independent.

It is shown how a multidimensional trait can be introduced into the complexity factors model, thus weakening Assumption 1. The factorization of the likelihood made possible by Assumption 2 shows the real need for strategy application traits that are not as complicated to estimate, as indicated by the Embretson paper. In addition, the factorization of the likelihood implies plausible models for the ICC of the total item response, incorporating parameters for guessing and sloppy responses. The ICC of the total item response takes on a very limited, multiplicative formula incorporating a multidimensional trait under Assumption 3. However, this limited ICC is not unreasonable for traits described in the cognitive processes formulation.

References

Barton, M. & Lord, F. M. An upper asymptote for the three-parameter logistic item-response model (Research Report RR-81-20). Princeton NJ: Educational Testing Service, 1981. (Unpublished)

Lazarsfeld, P. F., & Henry, N. W. Latent structure analysis. Boston: Houghton-Mifflin, 1968.

Acknowledgment

This research was supported by the Personnel and Training Research Programs, Office of Naval Research, under Contract Number N00014-83-C-0627, Contract Authorization Identification Number NR150-522, Douglas H. Jones, Principal Investigator, Advanced Statistical Technologies Corporation, 10 Trafalgar Court, Lawrenceville NJ 08648.

A LATENT TRAIT MODEL FOR INTERPRETING MISCONCEPTIONS IN PROCEDURAL DOMAINS

KIKUMI K. TATSUOKA
UNIVERSITY OF ILLINOIS

The widespread availability of computers has brought a new dimension to the theory and practice of educational and psychological testing in the past several years. A new testing procedure that is not linear in form, adaptive testing, has been introduced and has demonstrated its superiority to conventional linear testing. Item response theory has proved to be a useful technique to measure the performance on such an individualized test, with scores obtained from the computerized adaptive testing procedure.

Another development remarkably different from the traditional psychometric method is that of error diagnostic testing, which can determine the sources of misconceptions as they occur along the steps necessary to reach the correct answer to a given problem. These approaches--fully utilizing the capability of computers--have successfully diagnosed hundreds of erroneous rules resulting from a variety of misconceptions, or incomplete knowledge, of one or two procedural steps in domains such as the arithmetic of whole numbers, signed numbers, and algebra (Brown & Burton, 1978; Matz, 1980; Shaw, Standiford, Klein, & Tatsuoka, 1982; Tatsuoka, Birenbaum, Tatsuoka, & Baillie, 1980; van Lehn, 1982).

It is a well-known fact among cognitive psychologists that correct answers can often be obtained for wrong reasons. Tatsuoka and her associates (Birenbaum & Tatsuoka, 1983; Tatsuoka & Tatsuoka, in press) have demonstrated that measuring students' performances by considering the total number of correct answers to be the score can be a serious mistake for achievement tests that are especially designed to measure the outcome of learning. According to Davis and McKnight (1980), Tatsuoka and Tatsuoka (1982), and van Lehn (1982), erroneous rules can produce the correct answer for a given item. For example, the correct answer to the problem -16 - (-4) can be obtained by the following three erroneous rules: (1) subtracting the two numbers and taking the sign of the number with the larger absolute value; (2) changing the minus operation sign to addition, misunderstanding the parenthesis as the bars for absolute value and then applying the correct rule for addition; or (3) converting the subtraction to an addition problem by changing the sign of the second number, then subtracting the smaller absolute value from the larger absolute value and taking the sign of the first number for the answer. These three erroneous rules, which are committed by a substantial number of seventh graders, produce the correct answer (-12). Indeed, the third rule just described can produce the correct answers for all subtraction problems in which the first number has a larger absolute value than the second number.

Obtaining the correct answers for the wrong reasons reduces the reliability of the test scores and leads to multidimensionality of data sets for achievement tests of the problem-solving type (Birenbaum & Tatsuoka, 1982; Tatsuoka & Birenbaum, 1981). The relationship between such psychometric properties and misconceptions has been investigated by Tatsuoka and her colleagues (Birenbaum & Tatsuoka, 1983; Tatsuoka & Tatsuoka, 1981, 1982a). This series of studies shows that when achievement tests are used as an integral part of instruction, performance on the tests indicate how well students master the use of the correct rule (taught by instruction) for responding to the items. If the data contain many responses resulting from consistent application of erroneous rules, then the dimensionality (in the factor analytic sense) of the data becomes larger. These studies imply that it is impossible to separate cognitive aspects of learning from psychometrics in order to interpret properly the students' performances on achievement tests when the tests are used to measure outcomes of instructional treatment.

Purpose

This study will introduce a measurement model using item response theory for dealing with the misconceptions committed by many students. It is useful to know the transitional behavior of error types which may be due to a change of instructional methods, to advancement of learning stages, or to the stability and persistance of particular misconceptions. Such knowlege helps to evaluate instruction, to measure the outcome of learning, and to obtain diagnostic information for designing remedial instruction depending on the types of misconceptions. The purpose of the model is to express such functional aspects of misconceptions quantitatively so that they can be related to other measures such as motivation or creativity.

Finally, a technique will be discussed for assessing erroneous rules when they are used inconsistently by a student due either to "slips" or to the instability of the students' misconceptions. The pattern classification technique to determine the student's latent state of knowledge or the underlying misconception seems very useful in handling the variability of errors.

Rule Space

Tatsuoka and Baillie (1982b) showed that all erroneous rules of operation in signed-number arithmetic were expressible as points in a geometric space called "rule space." In other words, rule space is a geometric representation of the rules used. Brown and Burton (1978), van Lehn (1982), Tatsuoka and Tatsuoka (1981), Birenbaum and Tatsuoka (1980), and Matz (1980) have found numbers of erroneous rules resulting from a variety of misconceptions in whole number subtraction problems, signed-number addition and subtraction problems, and algebra. Those rules are identified by closely examining responses obtained from a set of items that are carefully chosen so as to enable distinguishing each rule from others (Tatsuoka et al., 1980; van Lehn, 1982). For example, all erroneous rules of operation in signed-number addition and subtraction problems which have been discovered so far are represented by a unique set of responses to the set of items. Tatsuoka and Linn (1981) mapped a set of binary-score vectors of these responses into the space spanned by the total scores and the values of the extended caution index, ECI4 (see below), along with the response vectors generated by 15 erroneous rules. In this space (referred to hereafter as the rule space), the erroneous rules resulting from the same kind of misconceptions clustered closely. This finding was confirmed in the domain of fraction addition problems (Tatsuoka & Chevalaz, 1983). Because of the nature of the extended caution index, unusual responses to the items are located in the lower or upper parts of the space and very usual responses scatter closely along the x-axis (the total scores or true scores).

Extended Caution Indices

A group of extended caution indices was introduced by Tatsuoka and Linn (1983), and their statistical properties were investigated in Tatsuoka and Tatsuoka (1982b). The values of the ECIs are calculated by first constructing two matrices; one is a binary score matrix

$$(y_{ij}),\ i = 1, \ldots, N,\ j = 1, \ldots, n \qquad [1]$$

where N is the number of students and n is the number of items in a test. The other is a probability matrix with elements P_{ij}, where P_{ij} is a logistic function of one, two, or three parameters.

In practice, the estimated P_{ij} obtained by substituting estimated item and person parameters in the logistic function can be used. One of the ECIs, ECI4, is defined as an index reflecting anomaly of an actual response pattern at a given level of ability θ_i. It is given by the complement of the ratio of two covariances: the numerator is the covariance of the ith row vector, y_i, in (y_{ij}) and the ith row, P_i, of the probability matrix (P_{ij}); the denominator is the covariance of the column-mean vector of $\underset{\sim}{G} = (G_{.1}, G_{.2}, \ldots, G_{.n})$ and the ith row vector P_i, both of (P_{ij}). Here

$$G_{\cdot j} = \frac{1}{N} \sum_{i=1}^{N} P_{ij} \ . \qquad [2]$$

That is,

$$\mathrm{ECI4} = 1 - \frac{\mathrm{cov}(\underset{\sim}{y}_i\ ,\ \underset{\sim}{P}_i)}{\mathrm{cov}(\underset{\sim}{G}\ ,\ \underset{\sim}{P}_i)} \ . \qquad [3]$$

The conditional expectation and variance of ECI4 are given by

$$E(\mathrm{ECI4}|\theta_i) = 1 - \frac{\mathrm{Var}(\underset{\sim}{P}_i)}{\mathrm{cov}(\underset{\sim}{G}\ ,\ \underset{\sim}{P}_i)} \ , \qquad [4]$$

$$\mathrm{Var}(\mathrm{ECI4}|\theta_i) = \frac{\sum_{j=1}^{n} \sigma_{ij}^{2}(P_{ij} - T_i)^2}{n^2 \mathrm{cov}^2(\underset{\sim}{G}, \underset{\sim}{P}_i)} . \qquad [5]$$

Thus, the standardized ECI4 is given by

$$\mathrm{ECI4}_z = \frac{n[\mathrm{cov}(\underset{\sim}{P}_i - \underset{\sim}{y}_i, \underset{\sim}{P}_i)]}{\sum_{j=1}^{n} \sigma_{ij}^{2}(P_{ij} - T_i)^2} \qquad [6]$$

where

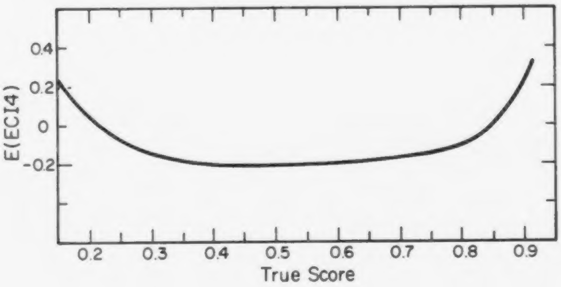
$$T_i = \frac{1}{n} \sum_{j=1}^{n} P_{ij} \qquad [7]$$

the raw-mean vector of (P_{ij}) and

$$\sigma_{ij}^{2} = P_{ij}(1 - P_{ij}), \qquad [8]$$

the variance of item _j_ at level _i_.

Figure 1
Expectation of ECI4 Against True Score



The figures of the expectations obtained from two different data sets, one being from a multiple-choice test and the other from a free-response item test, show an interesting difference. The expectation of the NAEP data set (a 68-item mathematics test used in the National Assessment of Educational Progress) shows a U-shaped curve (Figure 1), while that of the free-response test has a monotonically increasing curve (Figure 2). Figure 3 is the curve of the conditional standard error of ECI4 obtained from the same NAEP data set of 2,400 subjects. Tatsuoka and Tatsuoka (1982b) showed empirically that the standardized ECI4 has a normal distribution. It is not surprising because ECI4 is a weighted arithmetic mean of P_{ij}, $j = 1, 2, \ldots, n$.

Figure 2
Expectation of ECI4 Plotted Against the True Score

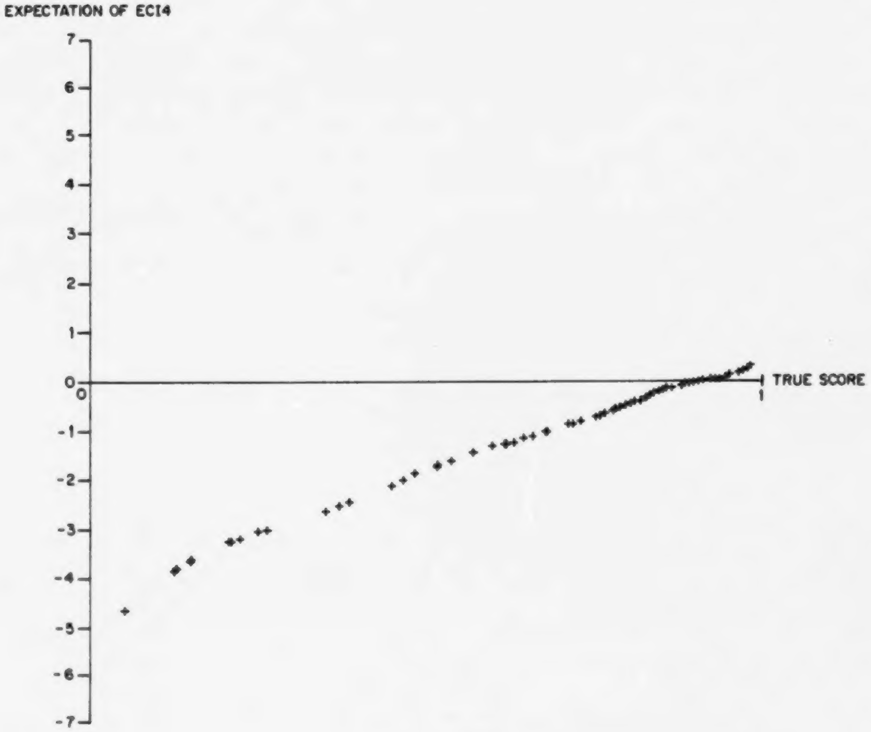
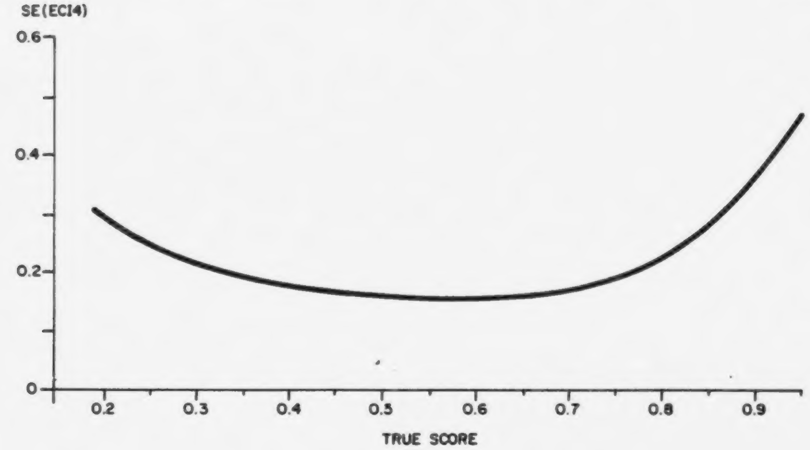


Figure 3
Standard Error of ECI4 for NAEP Data



A Map of Regions Classified by Misconceptions

One of the desirable properties of a rule space is that the responses resulting from the same kind of misconceptions are mapped into the neighborhood of the same kinds (Tatsuoka & Chevalaz, 1983; Tatsuoka & Linn, 1983). Tatsuoka and Chevalaz (1983) generated simulation data by controlling various sources of misconceptions in fraction addition problems and mapping them into the rule space. At the same time the responses generated by 52 erroneous rules described in Shaw, et al. (1982) were mapped into the same space along with the simulated responses. Figure 4 shows a map of those misconceptions.

Figure 4
Plot of True Score for Fraction Addition Test Against
Standardized Extended Index $ECI4_z$

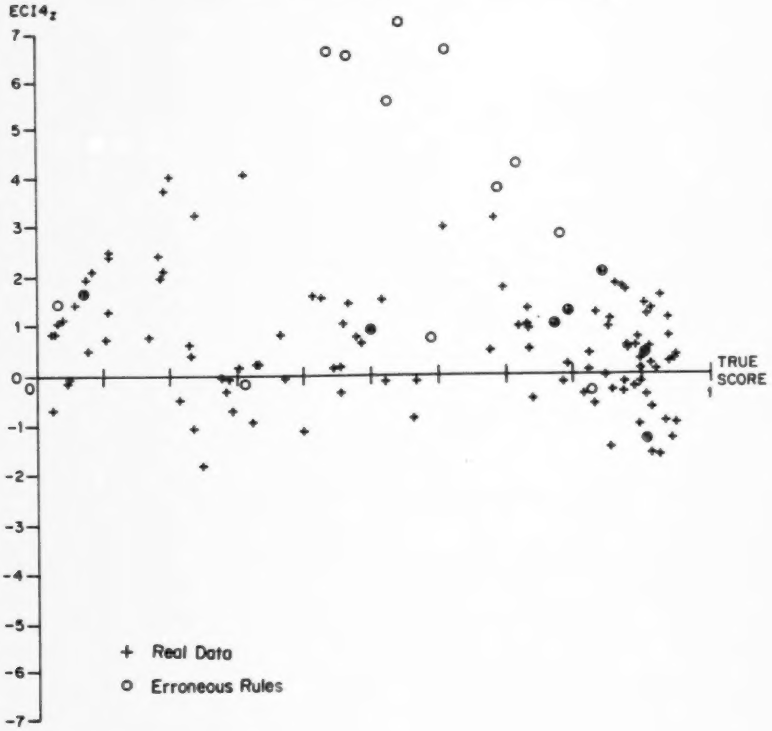


Table 1 describes the regions of A, B, ..., L in detail. For example, the response patterns (scored by the traditional scoring procedure) of Rules 39 and 10 fall in Region K, which is in the neighborhood of the responses consisting of correct answers for all mixed fraction problems and incorrect answers for others (or fraction-plus-fraction type). Rules 39 and 10 follows:

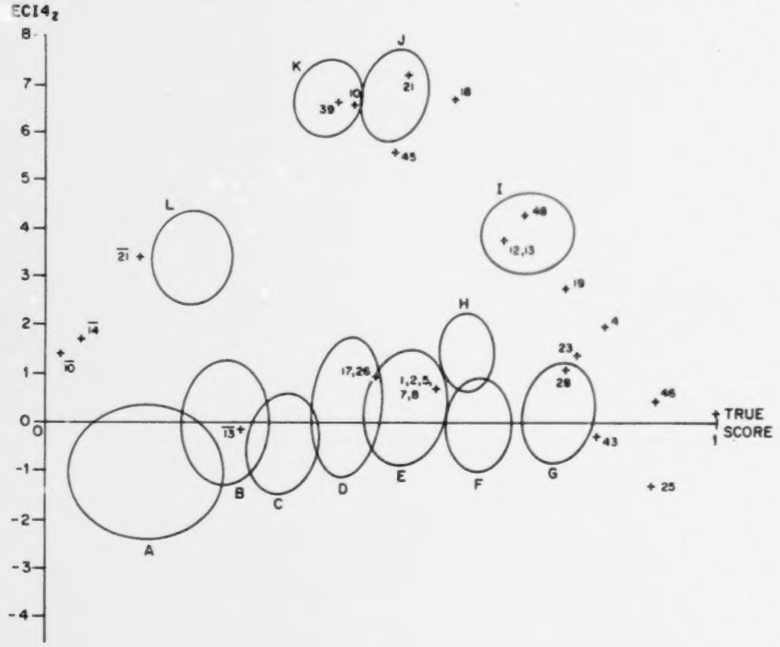
Rule 39: If the item was F+F type, then the numerator and denominator of the first fraction were added to obtain the numerator and the same was done for the second fraction to obtain the denominator. If the item was a mixed type, then the correct procedure was carried out.

BEST COPY AVAILABLE

Rule 10: A "1" was appended to all fractions (not mixed type), and then the resulting fraction was converted to an improper fraction. The correct procedure was then applied.

Rules 39 and 10 produce the correct answers for mixed types but always yield incorrect answers for F+F types. These responses result from some unusual erroneous rules; thus, the corresponding points are located at the upper middle section of the space in Figure 5.

Figure 5
Regions Determined by Various Misconceptions in the Fraction Addition Test and Erroneous Rules Described in Shaw et al. (1982)



On the other hand, the responses yielded by Rule 1 are correct answers for F+F types but incorrect for all mixed type items. Rule 1 is as follows:

Rule 1: The whole number was multiplied by the denominator and then the numerator was added. The denominator was not written, and the result was a whole number.

Figure 5 shows that the point in the rule space corresponding to Rule 1 belongs to Region E, whose points were generated as neighboring points of the responses of $x_j = 1$ if item $\underline{j}$ is an F+F type and $x_j = 0$ if otherwise.

It is interesting to note that the space in Figure 5 can be divided into several regions, each representing a unique source of misconception. In order to compare the regions in Figure 5 with actual students' responses from the fraction test, 186 students' response patterns are mapped into the space spanned by true scores and ECI4. Figure 4 shows their plots. By superimposing Figure 4 onto Figure 5, a student's underlying misconception, if there is one, can be

Table 1
Description of the 12 Regions Appearing in Figure 5

Region	No. of Items	Description of Region
A	8	If item _j_ is of F+F type with equal denominators or if one denominator is divisible by the other, and the numerator of the answer is smaller than the denominator, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
B	12	If _j_ is of F+F type and the denominators are equal, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
C	12	If j is of F+F type and the numerator of the answer is smaller than the denominator, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
D	22	If the denominators are equal, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
E	22	If _j_ is of F+F type then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
F	20	If the answer needs to be converted, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
G	32	If _j_ is of F+F type or has equal denominators, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
H	34	If _j_ is of F+F type or has different denominators, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
I	10	If _j_ is of mixed type and has equal denominators, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
J	18	If _j_ has different denominators, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
K	10	If _j_ is of mixed type, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.
L	10	If _j_ is of F+F type and has different denominators, then $x_{ij} = 1$; otherwise, $x_{ij} = 0$.

diagnosed. If the student's response falls in Region B, then the description given in Table 1 provides the information about how the student responded to the items. However, not all responses fall in one of the regions without any ambiguity. Some responses may belong to a gray area between two regions, and it is difficult to describe specifically which region of the misconceptions is really the student's source of errors. It is necessary to apply the pattern classification technique and to establish the boundary lines between two neighboring regions.

The numbered points in Figures 4 and 5 corresponding to erroneous rules of operation in fraction addition problems (see Shaw, et al., 1982) are often marked by two or more different numbers together. For example, Rules 1, 2, 5, 7, and 8 are represented by a single point. Rules 17 and 26 also share a single point. Since the true scores (x-axis) and the values of $ECI4_z$ (y-axis) are calculated from the binary response patterns obtained by the traditional "correct or incorrect" scoring procedure, these rules happened to produce the identical binary response patterns. In order to separate two different erroneous rules into different points in a geometric space, Tatsuoka and Baillie (1982) have proposed a new scoring procedure.

Component Response Patterns: Decomposing the Regular Scoring Procedure into Finer Components

Tatsuoka and Baillie (1982) demonstrated that several erroneous rules whose response patterns are identical with the traditional scoring procedure can be distinguished by decomposing the unit of the answer into finer components. Tatsuoka and Tatsuoka (1981) listed the response patterns of 45 erroneous rules in signed-number arithmetic also obtained from the regular scoring procedure. Some of the 45 binary patterns of 16 items were identical, although the descriptions of the erroneous rules which produced these identical patterns were not. There is no way to distinguish two such different rules just by examination of their binary response patterns.

However, if the regular scoring procedure is decomposed into finer components (e.g., the sign part of the answer for a given item and the absolute value part of the answer in signed-number problems or the whole number, numerator, and denominator parts of the answer for a fraction problem), then all the erroneous rules that have been discovered so far of the problem types in both signed-number and fraction arithmetic can be expressed uniquely as sets of the binary response patterns obtained from the component scoring procedure. For example, the response patterns of Rules 16, 32, 12, and 46 are given by:

Rule 16: The student subtracts the smaller absolute values from the larger absolute value and takes the sign of the number with the larger absolute value in his/her answer.
Rule 32: The two numbers are always subtracted as seen in Rule 16 but the "+" sign is always taken in the answers.
Rule 12: The student converts −2 − 8 and −13 − 5 into −2 + 8 and −13 + 5, respectively, but the other eight items are converted to addition problems correctly. Then the right addition rule is used to answer them.
Rule 46: The student has a strange idea about the parentheses and converts operation sign "−" to "+" first. Then he/she follows the rule: if the sign of the two numbers are minus, then change the sign of the second number to a "+"; if the signs of the two numbers are not alike, then the sign of the second number becomes a minus.

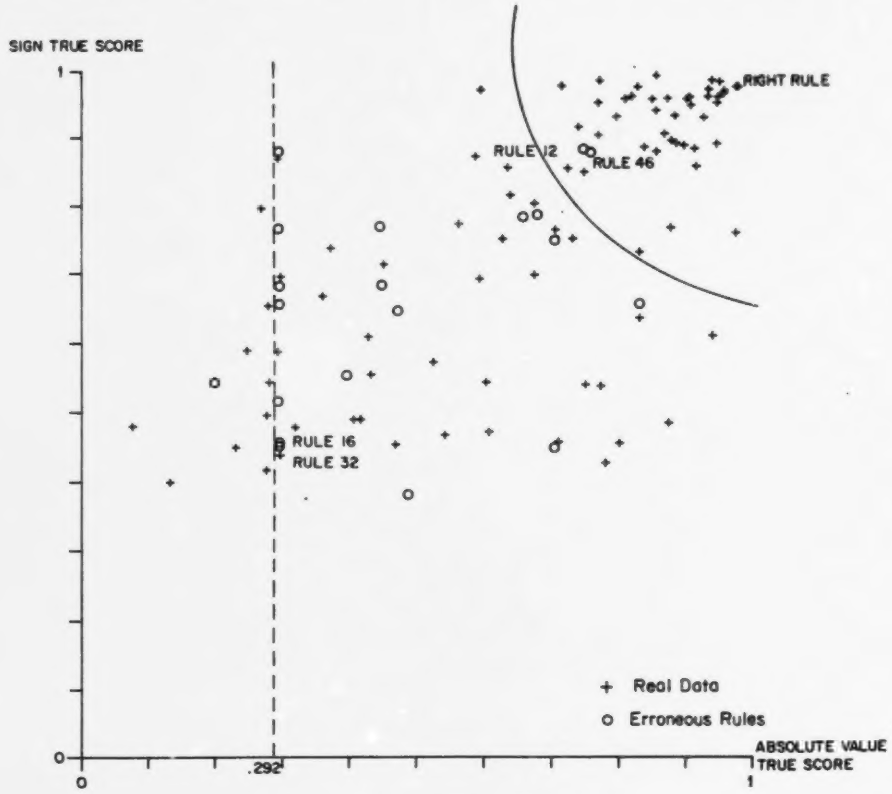
Table 2 shows that the regular response patterns for these four rules are elementwise products of the two component response patterns.

This new scoring method, the component scoring procedure, has been used in this study. Although the rationale of components scoring is based on a signed-number study, it can be generalized to other testing domains. For example,

Table 2
The Binary Response Patterns Generated by Four Rules and Three Different Scorings (R = Regular Scores, S = Sign-Component Scores, and A = Absolute-Value Component Scores)

	Rule 16				Rule 32				Rule 12				Rule 46			
Items	Response	R	S	A	Response	R	S	A	Response	R	S	A	Response	R	S	A
-3 - (-7) +4	-4	0	0	1	+4	1	1	1	+4	1	1	1	+4	1	1	1
-2 -8 = -10	+6	0	0	0	+6	0	0	0	+6	0	0	0	-10	1	1	1
5 - (-12) +17	-7	0	0	0	+7	0	1	0	+17	1	1	1	-7	0	0	0
-11 - +8 -19	-3	0	1	0	+3	0	0	0	-19	1	1	1	-19	1	1	1
9 - 4 = +5	+5	1	1	1	+5	1	1	1	+5	1	1	1	+5	1	1	1
-15 - (-9) = -6	-6	1	1	1	+6	0	0	1	-6	1	1	1	-6	1	1	1
-13 - 5 = -18	-8	0	1	0	+8	0	0	0	-8	0	1	0	-18	1	1	1
8 - (-6) +14	+2	0	1	0	+2	0	1	0	+14	1	1	1	+2	0	1	0
-5 - +11 -16	+6	0	0	0	+6	0	0	0	-16	1	1	1	-16	1	1	1
1 - 10 = -9	+9	0	0	1	+9	0	0	1	-9	1	1	1	-9	1	1	1

Figure 6
Plot of True Score for Absolute Value Component Against Sign Component



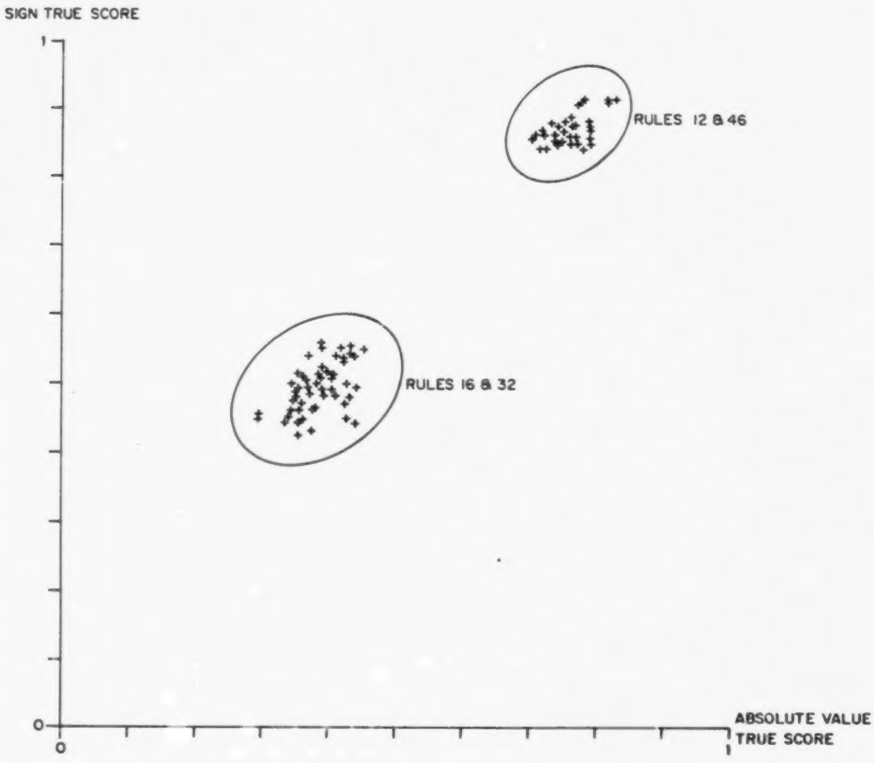
fraction problems can be decomposed into a whole number, numerator, and denominator parts.

Various Rule Spaces

A 40-item free-response test comprised of four parallel subtests of signed-number subtraction problems was administered to 172 eighth graders at a local junior high school. The traditional scoring of correct or incorrect answers was decomposed into a two-component scoring procedure of the absolute value and sign parts of the responses. The signs of the responses to the 40 items were scored correct or incorrect and so were the absolute values. The two-component response patterns were separately subjected to the estimation of item and person parameters of the 2-parameter logistic model. As was mentioned earlier, a rule space is defined as a geometric representation of the rules (including the correct rule and inconsistent application of two or more rules) used by the students.

The 20 small circles in Figure 6 represent 20 different erroneous rules, while the plus signs stand for a set of real responses to the 40 items for a

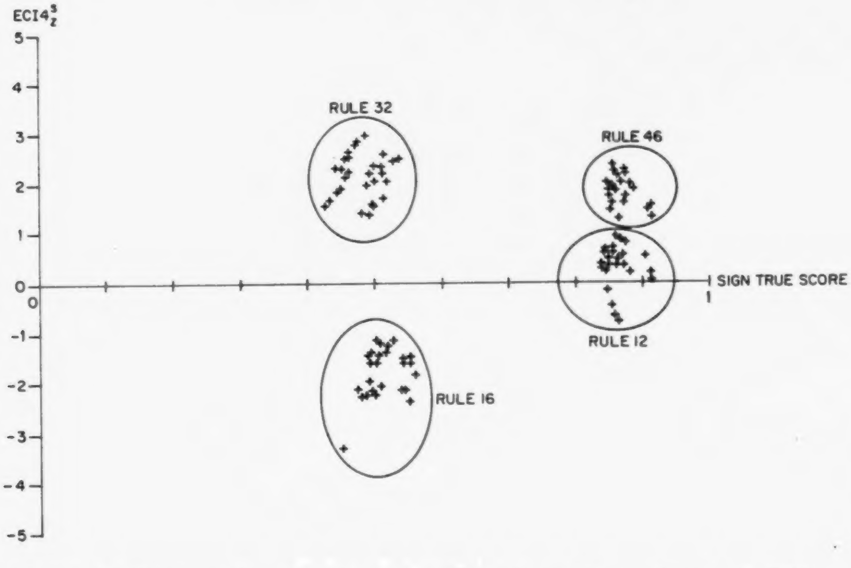
Figure 7
Plot of the Sign True Score Against Absolute Value for the Four Clusters Around Rules 12, 16, 32, and 46 in Figure 5



student. If the student responded to the 40 items by applying his/her erroneous rule consistently, then the student's point should coincide with the circle representing the rule. Figure 6 shows two such points. Most points do not show overlap; however, some real responses are located in the vicinity of a rule.

Tatsuoka and Baillie (1982) generated simulation data of responses resulting from inconsistent application of a rule. One or two items out of the 40 items do not follow a given erroneous rule and thus the component response patterns do not match completely with the patterns produced by the erroneous rule. Based on the 20 different rules, 20 sets of simulation data were generated and plotted on the space spanned by both the component true scores. As can be seen in Figure 7, Rules 16, 32, 12, and 46 and their simulated responses are not separated well because Rules 16 and 32, and 12 and 46, respectively, are already represented by very close circles in Figure 6. When plotted in terms of the sign true score against the standardized ECI4 obtained from the sign-component scores, four distinctly different clusters are formed in the new rule space, as shown in Figure 8. It is apparent that the values of ECIs are capable of separating response patterns that have either very close true scores or the same total scores.

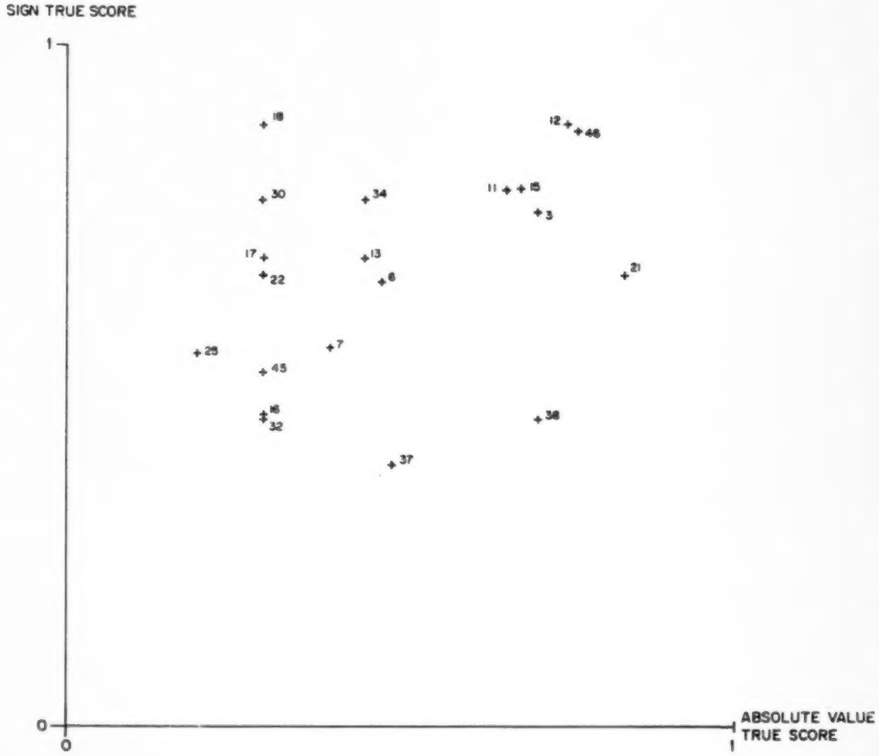
Figure 8
Plot of the Clusters Around Rules 12, 16, 32, and 46 in Figure 5,
Sign True Score Against $\mathrm{ECI4}^{s}_{z}$



Pattern Classification

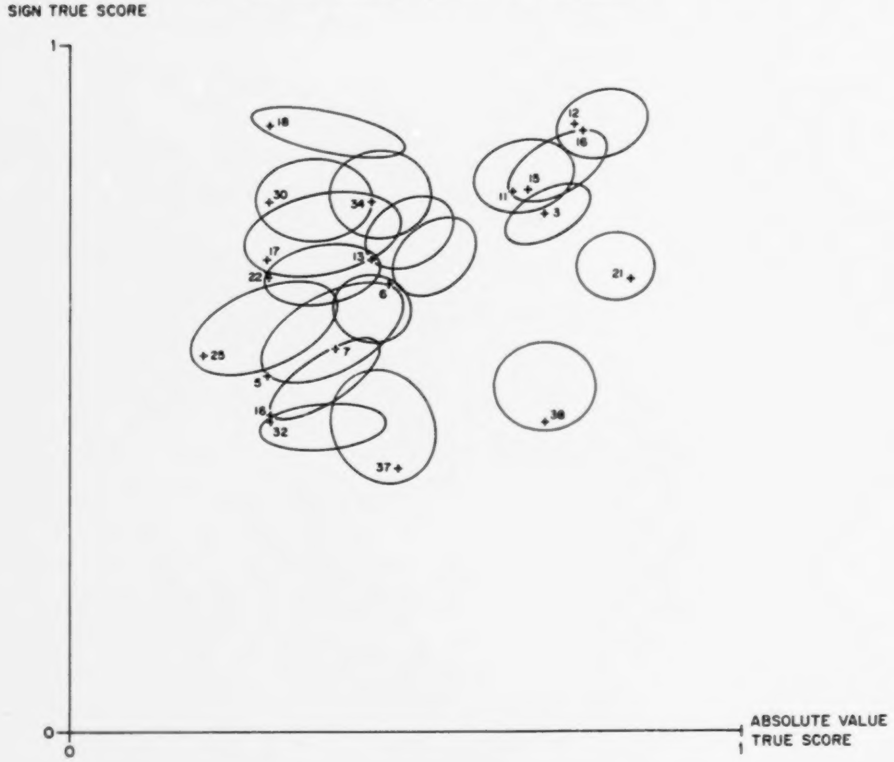
In the previous section, Figure 8 showed the four erroneous rules; the non-consistent responses neighboring each of them form four distinctly different clusters. By calculating a set of linear classification functions of the four clusters and by setting the boundaries to divide the four regions, it is possible to identify the underlying misconception of a new response with some probability of error by examining in which region the new response falls. This is a procedure for pattern classification and recognition problems (Fukunaga, 1972). Thus, the problem of diagnosing an individual student's misconceptions has finally been transmuted into a classification problem. There are, however, many more erroneous rules than just four to which the classification problem can be applied, even in a simple arithmetic domain. Brown and Burton (1978) have discovered approximately 200 misconceptions, Tatusoka, et al. (1980), have found at least 30 erroneous rules in signed-number addition problems, and there are more erroneous rules in signed-number subtraction problems. A computer program called SIGNBUG has been developed for diagnosing these hundreds of erroneous rules in signed-number arithmetic tests (Tatsuoka & Baillie, 1982a). The logic of the mechanism is deterministic, however, and if a student responds to an item without using his/her rule, then SIGNBUG cannot determine the rule. A revised version of SIGNBUG can now diagnose a partial application of the student's rule to the test items, as when the student uses the correct rule for a subset of the test items and uses some erroneous rule for the rest of the items; but it is difficult to cope with responses yielded by random "slips." With the probabilistic approach of rule space and pattern classification, it is possible to remedy the weakness of the deterministic approach without losing the strength provided by SIGNBUG.

Figure 9
Twenty Erroneous Rules of Signed-Number Subtraction Problems



In general, pattern classification techniques must be applied to distinguish the 20 or more erroneous rules from one another. As has been mentioned earlier regarding fraction addition problems, the responses resulting from the same (or similar) kinds of conceptual mistakes tend to cluster closely. Moreover, the signed-number subtraction test demonstrated (Tatsuoka & Linn, 1983; Tatsuoka & Tatsuoka, 1981) a result very similar to the phenomenon seen in Figure 5. As a result, the pattern classification technique may be suitable for determining the source of misconceptions instead of considering a finer object--an erroneous rule of operation and its neighboring nonconsistent responses.

Figure 10
Twenty Clusters of the Responses Neighboring
the 20 Rules Shown in Figure 9



In this study, the focus was on determining the classification boundaries of 20 clusters neighboring each of the 20 erroneous rules of signed-number subtraction problems. The list of the 20 rules and their code numbers are given elsewhere (Tatsuoka & Tatsuoka, 1981) and are represented by the plots in Figure 9. Figure 10 shows the 20 clusters around each of the 20 rules. A stepwise discriminant analysis (BMDP7M) was used to determine the classification functions, and four independent variables--absolute value and sign true scores

Table 3

Classification Matrix: Number of Cases Classified into Group and Percent Correct

Rule	Percent Correct	Rule																			
		3	6	7	11	12	13	15	16	17	18	21	22	25	30	32	34	37	38	45	46
3	100	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	100	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	100	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	100	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	100	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	100	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	100	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0
16	100	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0
17	100	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0
18	100	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0
21	100	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0
22	100	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0
25	100	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0
30	100	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0
32	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0
34	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0
37	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0
38	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0
45	96.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0
46	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31
Total	99.8	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31	31

and $ECI4_z$ for absolute value and signs--were used in the analysis. The result is summarized in Table 3. Nineteen of the rules were perfectly classified without any error of classification, and only Rule 45 has one out of the 31 samples misclassified.

Summary and Discussion

This study introduced a probabilistic model utilizing item response theory for dealing with a variety of misconceptions. The model can be used for evaluating the transition behavior of error types, advancement of learning stages, or the stability and persistence of particular misconceptions. Moreover, it can be used for relating the functional behaviors of errors to other criterion measures such as creativity, anxiety, and motivation.

One of several personal indices based on item response theory was used to formulate the "rule space," which is a geometric representation of erroneous rules of operation. ECI4, one of the indices, which is used primarily for detecting aberrant response patterns, proved to be effective for separating clusters of response patterns from one another.

A cluster of response patterns consists of the response pattern yielded by some rule and its "slips," due to partially consistent application of the rule. Using pattern classification to separate the clusters in the rule space accounts for variability of errors in the model. It thus seems to be a promising approach toward assessing an individual's state of knowledge.

REFERENCES

Birenbaum, M., & Tatsuoka, K. K. The use of information from wrong responses in measuring students' achievement (Research Report 80-1-ONR). Urbana: University of Illinois, Computer-based Education Research Laboratory, 1980. (NTIS No. AD A097715)

Birenbaum, M., & Tatsuoka, K. K. On the dimensionality of achievement test data. Journal of Educational Measurement, 1982, 19, 259-266.

Birenbaum, M., & Tatsuoka, K. K. The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 1983, 20, 17-26.

Brown, J. S., & Burton, R. R. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 1978, 2, 155-192.

Davis, R. B., & McKnight, C. The influence of semantic content on algorithmic behavior. The Journal of Mathematical Behavior, 1980, 3, 39-87.

Fukanaga, K. Introduction to statistical pattern recognition. New York: Academic Press, 1972.

Matz, M. Toward a computational theory of algebraic competence. The Journal of Mathematical Behavior, 1980, 93-166.

Shaw, D. J., Standiford, S. N., Klein, M. F., & Tatsuoka, K. K. Error analysis of fraction arithmetic: Selected case studies (Research Report 82-2-ONR). Urbana: University of Illinois, Computer-based Education Research Laboratory, February 1982.

Tatsuoka, K. K., & Baillie, R. SIGNBUG: An error diagnostic computer program for signed-number arithmetic on the PLATO® system, 1982. (a)

Tatsuoka, K., K., & Baillie, R. Rule space, the product space of two score components in signed-number subtraction: An approach to dealing with inconsistent use of erroneous rules (Research Report 82-3-ONR). Urbana: University of Illinois, Computer-based Education Research Laboratory, 1982. (b)

Tatsuoka, K. K., & Birenbaum, M. The effect of different instructional methods on achievement tests. Journal of Computer-based Instruction, 1981, 8, 1-8.

Tatsuoka, K. K., Birenbaum, M., Tatsuoka, M. M., & Baillie, R. A psychometric approach to error analysis on response patterns (Research Report 80-3-ONR). Urbana: University of Illinois, Computer-based Education Research Laboratory, 1980.

Tatsuoka, K. K., & Chevalaz, G. M. A map representation of misconceptions: An approach utilizing item response theory and classification functions (Research Report 83-4-ONR-NIE). Urbana: University of Illinois, Computer-based Education Research Laboratory, May 1983.

Tatsuoka, K. K., & Linn, R. L. Indices for detecting unusual response patterns: Links between two general approaches and potential applications (Research Report 81-5-ONR). Urbana: University of Illinois, Computer-based Education Laboratory, August 1981.

Tatsuoka, K. K., & Linn, R. L. Indices for detecting unusual response patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 1983, 7, 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. Spotting erroneous rules of operation by the individual consistency index (Research Report 81-4-ONR). Urbana: University of Illinois, Computer-based Education Research Laboratory, 1981.

Tatsuoka, K. K., & Tatsuoka, M. M. Detection of aberrant response patterns. Journal of Educational Statistics, 1982, 7, 215-231. (a)

Tatsuoka, K. K., & Tatsuoka, M. M. Standardized extended caution indices and comparison of their error detection rates (Research Report 82-4-ONR). Urbana: University of Illinois, Computer-based Education Research Laboratory, March 1982. (b)

Tatsuoka, K. K., & Tatsuoka, M. M. Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, in press.

van Lehn, K. Empirical studies of procedural flaws, impasses, and repairs in procedural skills (Technical Report ONR-8). Palo Alto CA: XEROX, Palo Alto Research Center, March 1982.

ACKNOWLEDGMENT

This research was sponsored by the Personnel and Training Research Program, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-79-C-0752. The authors gratefully acknowledge the assistance of Maurice Tatsuoka, Gerard Chevalaz, and John Eddins for editing and careful proofreading, and of Louise Brodie for typing and Roy Lipschultz for graphics. Some of the analyses in this report were done on the PLATO® System, which was developed by the University of Illinois. (PLATO® is a service mark of Control Data Corporation.)

DISCUSSION

SUSAN EMBRETSON (WHITELY)
UNIVERSITY OF KANSAS

The actual content of Tatsuoka's paper is broader than the title implies. Tatsuoka presents a general method for diagnosing qualitative differences in the strategies that are applied to solving test items. The approach is error diagnostic testing, in which the pattern of errors between items is used to diagnose the strategy (and conceptualizations) that the person applies to the problem.

In choosing arithmetic items to demonstrate the method, Tatsuoka has made a wise choice; children's misconceptions of number have been studied intensively in several studies (e.g., Resnick & Ford, 1981; Riley, Greeno, & Heller, in press) in addition to those mentioned by Tatsuoka. In general, misconceptions of number can lead to a correct answer in some problems. However, in certain problems misconceptions will consistently lead to errors. Tatsuoka labels the misconceptions that can be applied to arithmetic problems "erroneous rules."

The main aspect of the method is using two-dimensional space to locate regions that correspond to the various erroneous rules. The variables that define the space are ability (monotonically related to raw score on the items) and the extended caution index. The extended caution index assesses the fit of an individual item response vector to the pattern that would be expected for a person's ability level. For example, a low ability person would be expected to solve easy items and to answer incorrectly difficult items. Sometimes, however, a person's response pattern deviates from the expected pattern. Tatsuoka assumes that misfit is diagnostic of strategy differences.

The extended caution index that is reported in the Tatsuoka paper uses a latent trait model to generate the item probabilities expected from the person's ability level. Note that it is not necessary to have a latent trait model to generate these expectations (although it is the most theoretically adequate method). Expectations could be generated from item probability marginals, for example.

The goal of Tatsuoka's method is to classify persons according to the conception of number that best explains their performance. Regions for the various erroneous rules are determined by generating item response vectors to correspond to the pattern of correct and incorrect responses that would result from application of that rule. The extended caution index and "ability" are then computed from the response vector for the rule and located in the two-dimensional space. Several erroneous rules are thus located. Then, real data are projected into the space. The assumption is that the closer a person is to a point defined by an erroneous rule, the more likely the rule explains his or her responses.

In general, I find this work fascinating. It integrates recent findings in cognitive theory with psychological measurement. If the project is successful, it should provide useful measurements for cognitive interventions. For example, a teacher could benefit greatly by diagnosing the concepts of number for students, so as to be prepared to correct erroneous conceptualizations.

Several aspects of Tatsuoka's method need further research. First, there are no data to support the assumption that the more similar a person's ability and misfit are to an erroneous rule, the more likely the rule has been applied. It could be that the person uses another (unmeasured) rule that generates the very same response pattern on the item set. Perhaps the rule generates a different response pattern, but the same ability and misfit. Another possibility is that the person applies different rules to different tasks. Thus, the diagnosis yielded from error patterns needs to be confirmed with external data on strategies to assess the diagnostic potential of the method.

Second, the diagnostic potential of a particular item set for specified rules needs to be assessed. Two erroneous rules may be indistinguishable in one item set, yet quite distinct in another item set. Some type of discriminability index for the item set with respect to the specified strategies needs to be developed.

Third, the current paper really does not present a latent trait model for misconceptions and strategies. Instead, it uses a latent trait model in the context of another diagnostic method. I believe, however, that it may be useful to develop latent trait models indicating the probability that an individual is applying a specified strategy. I have examined this problem in my own research (Embretson, in press); and I believe that it would be feasible to develop similar models for misconception data, especially when within-item information is available. Tatsuoka presents some within-item data and gets fairly impressive categorization results.

REFERENCES

Embretson (Whitely), S. Latent trait model approaches to assessing and analyzing individual differences. In R. J. Sternberg (Ed.), *Approaches to individual differences*. Hillsdale, NJ: Erlbaum, in press.

Resnick, L. B., & Ford, W. *The psychology of mathematics for instruction*. Hillsdale, NJ: Erlbaum, 1981.

Riley, M. S., Greeno, J. G., & Heller, J. I. Development of chidren's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking*. New York: Academic Press, in press.

THE COMPUTERIZED ADAPTIVE TESTING SYSTEM DEVELOPMENT PROJECT

JAMES R. MCBRIDE AND J. B. SYMPSON
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

All four Armed Services use ASVAB (the Armed Services Vocational Aptitude Battery) to assess potential enlisted personnel, and to make decisions regarding personnel selection and classification. The Computerized Adaptive Testing (CAT) project is a joint-service coordinated effort to develop and evaluate a system for automated, adaptive administration of the ASVAB. If approved, the CAT system will be used by the Military Entrance Processing Command (MEPCOM) to conduct operational testing of applicants for military enlistment. Navy Personnel Research and Development Center (NPRDC) is the lead service laboratory for this research and development effort.

Project Background

A Description of CAT

CAT is a system for administering personnel tests. It differs from conventional test administration in two major respects:

1. Automated test administration using a computer terminal, rather than printed booklets and answer sheets;
2. Adaptive (tailored) sequencing of test questions, rather than the lock-step sequencing inherent in printed tests.

Automated test administration. In a CAT system, test questions are displayed on the video screen of a computer terminal. The examinee answers each question using an input device such as a keypad or a light pen. The test is automatically scored by a computer program.

Although a conventional computer terminal could be used for CAT, specially designed test administration terminals are preferable for several reasons, including: (1) the need for graphics display capabilities, since many questions include pictures and drawings; and (2) the fact that most examinees are unfamiliar with the use of computer terminals and typewriter-style keyboards.

Adaptive question sequencing. Examinees taking a conventional, printed test all answer the same questions, printed in the same sequence. Usually, easy questions are printed first and the more difficult ones are printed last. This lock-step sequencing wastes considerable amounts of time, because little useful information is gained by asking highly able examinees easy questions, or by challenging the less able examinees with questions far too difficult for them. From an aptitude measurement point of view, it is far more efficient to tailor the choice of test questions to the ability level of each individual examinee.

CAT does this. CAT makes successive approximations to the examinee's ability, updating the approximation after each question is answered. Each question in the test sequence is selected from a large bank of pre-calibrated questions to match the difficulty of the next question to the estimated ability of the examinee. This process of sequentially tailoring test difficulty to examinee ability is the essence of adaptive testing.

The current state-of-the-art in psychometrics will support the application of adaptive testing only to what are called "power" tests of cognitive aptitude; i.e., tests in which speed of the response is not a significant factor. Forms 8, 9, and 10 of ASVAB each contain eight power tests. Two other subtests (Numerical Operations and Coding Speed) are highly speeded; these speeded tests can be automated, but cannot presently be made adaptive.

Advantages of a CAT System

CAT has a number of potential advantages over conventional printed testing. Some of its advantages are due to automation; others are attributable to the adaptive nature of CAT.

1. Efficiency: CAT can reduce the length of many ASVAB tests by as much as 50%, without loss of measurement precision. This is a direct result of the adaptive sequencing of test questions.

2. Precision: ASVAB scores are precise only at mid-range ability levels. At the low and high extremes of ability, ASVAB is inherently imprecise. Because CAT matches test difficulty to the examinee's ability level, CAT scores will be substantially more precise than ASVAB in the extremes, and just as precise in the mid-range.

3. Accuracy: In the past, ASVAB tests were often manually scored. The raw test scores were transformed to standard scores, and then combined into aptitude-area composite scores, also manually. Finally, raw scores and composite scores were typed onto enlistment processing forms by clerks. All of these manual operations are susceptible to clerical errors, resulting in inaccurate data in the personnel record.

4. Security: Any printed or conventional test is susceptible to compromise. The probability of compromise increases directly with the frequency of administration and the duration of the operational life of a specific version of the test. In the past, ASVAB security violations have occurred; future compromise attempts seem inevitable as long as there is a significant incentive for applicants to perform well on military selection tests. CAT will eliminate the two major features that make printed ASVAB susceptible to compromise: pilferable test booklets, and predictable sequences of test questions. When implemented, the CAT system will also incorporate several

other features that will decrease the probability of test compromise.

5. Economy: The initial cost of creating the CAT system will be offset by substantial cost savings in several categories as the printed ASVAB is phased out. The cost of test administrators' labor should be significantly reduced, since CAT will take half as long as ASVAB to administer. Public burden costs (civilian test-takers' time) will be similarly reduced. Printing, distribution, and storage costs associated with the paper-and-pencil tests will be eliminated. And the cost of developing new forms of the tests will be substantially reduced.

6. Ease of Revision: Developing replacement forms of ASVAB has taken three to five years in the past, and has been very expensive and logistically cumbersome, due to the need for large-scale administration of experimental tests. In contrast, every administration of a CAT test can incorporate a small number of experimental test questions embedded in the operational test. These experimental questions will be unobtrusive, and will not interfere with the operational testing routine. This process will result in frequent periodic updating of the CAT item banks; i.e., revision of the tests will occur constantly, unobtrusively, and rapidly.

Psychometric Development of CAT

Psychometric methods and procedures to be employed in the CAT hardware/software system are being developed through a combination of contract and in-house research. Development is underway in the following areas:

1. Constructing calibrated item banks for the CAT system;
2. On-line calibration research;
3. Equating CAT with ASVAB subtests and service composites;
4. Validation of CAT as a measurement technique;
5. Evaluation of CAT's utility for predicting performance;
6. Meeting established professional standards for tests.

Constructing Calibrated Item Banks

Successful implementation of any CAT system requires the development of carefully calibrated item banks. Here, the term "calibration" refers to data analysis procedures that provide estimates of item response theory (IRT) parameters for each test question. IRT item parameters serve to describe the operating characteristics of test questions that are used to measure ability. The choice of which item to administer to an examinee at any point in an adaptive test is determined by a complex mathematical function of these parameters.

In connection with the joint-services CAT project, two major efforts have been undertaken in the area of item calibration. These may be identified as (1) experimental CAT system item bank development and (2) operational CAT system item bank development.

Experimental item bank development. In order to evaluate the reliability, construct validity, and criterion-related validity of CAT prior to its operational implementation, NPRDC has created an experimental CAT system. This small experimental system uses seven Apple III microcomputers, all connected to a Corvus hard-disk unit. The disk unit serves as a storage medium for the item banks, item parameters, and instruction files that are used by the experimental system. It also stores the data that are generated in connection with each examinee's testing session. At this time, development of the item banks needed for this experimental system is partially complete. Information about the calibrations that have been completed so far is presented in the paper following this one.

Operational item bank development. In order to have an adequate number of high-quality test questions available at the time that CAT is likely to become operational, Air Force Human Resources Laboratory (AFHRL) has contracted with Assessment Systems Corporation of St. Paul, MN, to develop and calibrate additional questions in the content areas associated with ASVAB "power" subtests. These questions, and possibly some of the better questions from the experimental item banks, will be used in the initial item banks for the operational CAT system. A target of 200 questions per adaptive-test item bank has been established.

On-Line Calibration Research

The two item-calibration efforts just described involve the calibration of test questions that have been administered in printed test booklets. There may be some risk in using item parameters obtained under this type of testing condition in the operational CAT system, since the medium for item presentation and examinee response will be quite different. An obvious solution for this potential problem would be to collect item calibration data on computer terminals. Unfortunately, item banks and subject samples as large as those required for the operational CAT system make this currently impossible. However, once the operational CAT system is in place, it will be a simple matter to insert two or three newly developed items into each CAT subtest that an examinee completes. Data collected in this manner (i.e., "on-line") could be used to calibrate new items very rapidly. Moreover, if research does indicate that the medium for test delivery has an influence on obtained parameter estimates, items appearing in the initial operational item banks could themselves be re-calibrated on-line.

At this time, well-developed statistical procedures for on-line item calibration do not exist. Some straightforward generalizations of current procedures could be implemented, but they probably would not be optimal. In order to foster the development of efficient on-line item calibration methods, and to make it likely that these methods will be available by the time that CAT becomes operational, the Office of Naval Research (ONR), NPRDC, AFHRL, and MEPCOM are planning to co-fund an ONR research contract in which several IRT researchers address the issues involved in on-line calibration. This research effort should result in one or more procedures that can be implemented in the operational CAT system.

Equating CAT with ASVAB

For military personnel selection and classification purposes, ASVAB subtest scores are combined into aptitude-area composites, such as the Mechanical, Administrative, General, and Electronics (M, A, G, and E) composites used by the Air Force. Each of the Services has its own composites and establishes composite qualifying scores for entry into specific occupational specialities. When CAT becomes operational, it will be essential to the continuity of the four Services' selection and assignment practices that CAT test scores be interchangeable with those of ASVAB. This will require that procedures be developed to "equate" CAT with ASVAB, i.e., to transform CAT scores onto the familiar ASVAB score scale so that a given score on a CAT test will have the same interpretive meaning and be useful for the same purposes as actual ASVAB scores. A committee of psychometric experts has been commissioned, with joint ONR/NPRDC funding, to develop methods for equating CAT to ASVAB.

Validating CAT as a Measurement Technique.

CAT is intended to measure the same aptitudes and abilities that ASVAB now measures. However, CAT is very different from ASVAB in its mode of administration and in other obvious respects. These obvious differences raise the possibility that CAT may measure something rather different from one or more of its ASVAB subtest counterparts. Consequently, it is necessary to conduct psychometric research to establish the construct validity of the CAT subtests. This research will involve administration of CAT and ASVAB tests to experimental groups of examinees, followed by statistical analyses to investigate the extent to which CAT and ASVAB measure the same aptitudes. The experimental CAT battery described above is currently being administered to military recruits in order to collect the data required for an assessment of CAT's construct validity.

Evaluating the Predictive Utility of CAT

ASVAB is useful as a tool for selection and classification of enlisted personnel by virtue of the known correlation of ASVAB subtest and composite scores with performance in occupational specialty training. Although there is every reason to think that CAT will be as useful as ASVAB for predicting training performance, it is still necessary to demonstrate CAT's predictive utility. To accomplish this, it is necessary to test recruits using CAT, and to conduct followup studies to determine the correlation between CAT test scores and training performance. Although it is not practical to do this for every occupational specialty in the four Services, it is essential to demonstrate that CAT has predictive utility across a broad spectrum of job types in each of the Services. Each of the four Services has been requested to identify six technical training courses to be involved in the demonstration of CAT's predictive utility. Recruits designated for subsequent assignment to those courses will be tested using the experimental CAT battery, and will be tracked through training in order to collect the criterion data required to evaluate CAT's predictive relationships with training performance data.

Meeting Established Professional Standards

Aptitude tests that are well designed and well developed can be enormously valuable to the institution that uses them, while at the same time being fair and equitable to examinees. Over the years, professional standards for the development and use of tests have been established. The best known of these are the "Standards for Educational and Psychological Tests," jointly published by the American Psychological Association, the American Educational Research Association, and the National Council for Measurement in Education. These standards are currently undergoing revision, with the revised standards due to be published in 1984.

It is essential that the CAT system comply with the highest professional standards for test use, in order to insure CAT's credibility as a valid, useful, and fair instrument for personnel assessment and selection. Meeting such standards will require that a variety of psychometric research studies be conducted and documented. The preferred method of documentation is the publication of a professional "test manual," which will contain all of the results of the research into CAT's psychometric characteristics and benefits.

Under an ONR contract, Bert Green of Johns Hopkins University has studied the current ASVAB test battery and the proposed CAT system, and has developed recommendations for evaluation of the CAT system; he was assisted in his efforts by psychometricians Darrell Bock, Lloyd Humphreys, Robert Linn, and Mark Reckase. The evaluation plan developed by this panel supplements the current APA/AERA/NCME Standards, for the special case of the CAT system. NPRDC is currently designing and conducting research to address the major points presented in the evaluation plan proposed by Green et al., and intends to report the results of this research in a test manual for the CAT system.

Criteria for CAT Delivery System Evaluation

The CAT system will become operational only after a milestone decision is made to implement it. That decision cannot be affirmative unless the CAT hardware/software system meets a number of criteria. Since the CAT system is being developed in several stages, intermediate evaluations can be conducted during the system development process, well in advance of the final evaluation that precedes the decision whether to implement the system. The following eight major criteria will be used in evaluating the CAT delivery system.

1. Performance. The CAT system will be evaluated as a computer system with a number of critical performance parameters. Included among these are:

   a. Speed of response: The system should respond to examinee input in less than 2 seconds. This is necessary in order to avoid distracting the examinee, and to ensure that the efficiency of adaptive testing will be reflected in reduced test administration time compared with ASVAB.

   b. Speed of display: Once the system has responded to examinee input, the next test question or dialogue frame should be completely displayed in less than 3 seconds, for the same reasons.

c. Display resolution: The display must provide clear, unambiguous presentation of both test and pictorial material. Alphanumerics should have a dot resolution of at least 7 × 9 for ease of reading. Pictorial material should have a resolution of at least 400 horizontal by 300 vertical picture elements for clarity.

d. Mass storage capacity: The system must provide adequate mass storage to contain system and applications programs, the large item banks used for adaptive testing, and archival records of each test administration and the test results. It should also have excess capacity, in order to be capable of growth in the number of subtests and in the size of the item banks.

e. Communications capability: The CAT system will be required to transmit large volumes of test results to a central computer site on a daily basis, and to receive software revisions and item bank updates periodically, using electronic data communications. In order to accomplish these functions economically and without interrupting the operational schedule of test administration, the system must be capable of high-speed data communications.

2. Suitability. The CAT system is intended for use in both Military Entrance Processing Stations (MEPS) and Mobile Examining Team sites (METs) without significant changes in the staffing or facilities currently available. This implies that CAT must be capable of operation by personnel of normal skill levels, without need for extensive specialized training. It must also be capable of operation in the normal office environment that characterizes MEPS and METs, without requiring special environmental controls on temperature and humidity. It should require a minimum of facility modifications. It should be portable if portability is required to serve MET sites.

3. Reliability and availability. The system must be available for testing whenever testing is scheduled; once testing has begun, all examinees must complete their testing. Little deviation from these requirements can be tolerated, due to the importance of timely completion of enlistment processing. Quantitative thresholds of 99.9% for both availability and reliability have been established as design goals. 99.9% availability means that a CAT installation would be unavailable for testing less than one day per 1,000 scheduled testing days, or about once every four years. 99.9% reliability means that less than 1 test per 1,000 would be interrupted by a system failure; redundancy features of the system will provide the capability to complete the few tests which are interrupted in this way.

4. Maintainability. The system must be designed so that there is no requirement for skilled technicians in the MEPS or METs. It should include built-in diagnostic tests to alert the test administrator of present or impending hardware or software failure. The system design should include integrated logistics support.

5. Ease of use. Because it will be used by examinees and test administrators with no computer experience, the system must be designed explicitly for ease of use. For the examinee, this means the system must use a very simple procedure for answering test questions, and must first teach the examinee how to use it. For the test administrator, the system must involve only a small number of simple operations, and must direct the test administrator as to what to do at each step in its use.

6. Security. The system must incorporate protection against unauthorized access to the CAT item banks, and to examinee records. It must make it impossible to make printed copies of test questions, and must contain no small pilferable articles whose loss could result in test compromise. It must make coaching ineffective as a means of test compromise, by eliminating predictable sequences of test questions. It must include a password access system, which not only prevents unauthorized access, but also creates an audit trail of every test administered on the system.

7. Affordability. A 10-year operating life is specified for the CAT equipment. The life-cycle cost of the CAT system over its 10-year life must be competitive with that of the printed ASVAB.

8. Flexibility. Much of the long-run potential of CAT resides in its capacity to support types of assessment that are either difficult or impossible to implement under paper-and-pencil modes of test administration. Some of these possibilities include answer-until-correct responding, open-ended responding, the use of dynamic task stimuli, and the assessment of perceptual and psychomotor skills. The new CAT system should have software/hardware capabilities that are sufficiently flexible to allow exploration and evaluation of such assessment capabilities.

Conclusion

Assuming the CAT system now under development is found to satisfy the many engineering and psychometric criteria that have been established as conditions for its acceptance, the phased implementation of CAT on a nationwide basis could begin as early as 1986. By 1990, over one million applicants for military service would be tested on the CAT system each year. This fact would undoubtedly provide a strong impetus for the application of computerized adaptive testing procedures in other areas of personnel selection and classification as well.

ITEM CALIBRATIONS FOR COMPUTERIZED ADAPTIVE TESTING (CAT) EXPERIMENTAL ITEM POOLS

J. B. SYMPSON AND LORALEE HARTMANN
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

In connection with the development of item pools for the Navy's experimental CAT system, items in five different content areas (General Science, Paragraph Comprehension, Word Knowledge, Math Knowledge, and Arithmetic Reasoning) have been calibrated using item response theory (IRT) methodology. Calibration of these items was originally undertaken by a government contractor (Research Applications, Inc.), but a review of the procedures followed by the contractor indicated that the calibrations should be repeated.

In the calibrations conducted at Navy Personnel Research and Development Center (NPRDC), operational Armed Services Vocational Aptitude Battery (ASVAB) subtests 8, 9, and 10 were calibrated along with the new computerized adaptive testing (CAT) items in order to insure that all estimated parameters in a content area were expressed relative to a common metric.

An example of the "linking design" for one content area (Arithmetic Reasoning) is shown in Table 1. Each row of this table (i.e., each "Group") represents approximately 312 male applicants for military service who completed one particular experimental Arithmetic Reasoning (AR) booklet and one particular operational ASVAB form. The experimental booklet and the ASVAB form completed by individuals in a given row of the table are indicated by Xs. Each experimental AR booklet contained 35 items and each ASVAB form contained 30 AR items. Thus, a total of 390 AR items are represented in the table. Each individual in the calibration sample was exposed to 65 of these items. An average of approximately 1,870 responses were available for each of the 390 AR items.

The number of experimental booklets, the number of CAT items, the number of ASVAB items, the total number of items, and the number of individuals in the final calibration sample for each content area are given in Table 2. As indicated by the column totals, the responses of 34,774 individuals were used to obtain IRT parameters for a total of 1,750 test items.

The computer program LOGIST, developed by Lord at Educational Testing Service, was used to fit the 3-parameter logistic model to the available item response data. Version 2.B of this program (released in 1976) was used, but several modifications to the program were made prior to completing the calibrations. These modifications were tested at NPRDC and found to improve the param-

Table 1
Linking Design for Arithmetic Reasoning Items

Group	CAT Booklet						ASVAB Form					
	B1	B2	B3	B4	B5	B6	8A	8B	9A	9B	10A	10B
1	X						X					
2	X							X				
3	X								X			
4	X									X		
5	X										X	
6	X											X
7		X					X					
8		X						X				
9		X							X			
10		X								X		
11		X									X	
12		X										X
13			X				X					
14			X					X				
15			X						X			
16			X							X		
17			X								X	
18			X									X
19				X			X					
20				X				X				
21				X					X			
22				X						X		
23				X							X	
24				X								X
25					X		X					
26					X			X				
27					X				X			
28					X					X		
29					X						X	
30					X							X
31						X	X					
32						X		X				
33						X			X			
34						X				X		
35						X					X	
36						X						X

eter estimates generated by LOGIST. In each content area, all CAT items and ASVAB items were calibrated simultaneously in a single LOGIST run.

Table 2
Characteristics of Calibration Database

Content Area	No. of CAT Booklets	CAT Items	ASVAB Items	Total Items	Sample Size
GS	2	200	150	350	4106
PC	4	180	90	270	7734
WK	2	200	210	410	4225
MK	4	180	150	330	7491
AR	6	210	180	390	11218
Total	18	970	780	1750	34774

Table 3 shows median values of the estimated IRT parameters $\hat{a}$, $\hat{b}$, and $\hat{c}$ that were obtained in each content area, separately for CAT items and operational ASVAB items. The median $\hat{a}$ values in column 1 of this table are high enough, and the median $\hat{c}$ values in column 3 are low enough, to suggest that CAT procedures will function quite effectively in each of these five content areas.

Table 3
Median Values of Estimated IRT Parameters

Content Area	CAT Items			ASVAB Items		
	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{a}$	$\hat{b}$	$\hat{c}$
GS	1.29	.02	.22	1.05	-.05	.22
PC	1.22	-.06	.23	.87	-.56	.23
WK	1.41	-.74	.25	1.45	-.15	.25
MK	1.38	.66	.18	1.66	.69	.19
AR	1.39	-.01	.22	1.52	.28	.22

During the next few months, items in four more content areas (Automotive Information, Electronics Information, Mechanical Comprehension, and Shop Information) will be calibrated using the procedures described above. Additional items will also be calibrated in order to augment three of the previously developed item banks.

COMPUTERIZED TESTING IN THE GERMAN FEDERAL ARMED FORCES (FAF): EMPIRICAL APPROACHES

WOLFGANG WILDGRUBE
PSYCHOLOGICAL SERVICE OF THE FAF

In April 1982, the Federal Armed Forces initiated the first empirical pilot project in the area of computerized adaptive testing (CAT) to collect data for psychological testing. Four microprocessor-based stand-alone computers called TEST 2000 (manufactured by the German firm Zak, which specializes in psychological and physiological measurement) were installed at the recruiting center for volunteers in Munich. This was the first installation for computerized testing in Germany and, according to present information, in Europe. TEST 2000 had previously been installed at some locations in Western Europe only for physiological measurement and in connection with an optical mark reader for preprocessing of data and simple analyses.

The first empirical phase started with a conventional testing strategy and the standard entrance test battery to collect data (1) for the examination of technical aspects (e.g., at the moment there are some problems with the keyboard for the examinees, and some trouble with the memory, which has become overheated during a period of hot weather); (2) for organizational aspects (computerized testing carried out during the entire selection process) and attitudinal change (e.g., reducing the psychologists' "computer anxiety"); and (3) for "human factor" aspects (e.g., motivation of volunteers, problems in handling the test computer). Collection of these first empirical data also presented an opportunity to compare paper-and-pencil testing versus computerized testing.

Configuration

Hardware

The testing station in Munich consists of four TEST 2000 computers. Each microprocessor-based stand-alone computer has a screen, processor, two floppy disks (a system disk and a data disk for approximately 80 persons), a special keyboard for the examinee (digits 1 to 9 and 0; four buttons: "correct," "incorrect," "I don't know," and "next item"), and a typewriter keyboard for the psychologist/proctor. It is possible to present the items in black on a green background or in green on a black background. One testing computer is connected with a line printer for the output of the testing results (The data disks from the three other computers are transferred to this computer, which is connected with the printer subsequent to the testing session).

TEST 2000 is based on a highly integrated unit of the ZILOG Z80A-family.

The stand-alone version offers the following:

1. 64 KByte RAM, 8KByte used by the operating system,
2. 8 KByte PROM containing integer- and floating-point arithmetic and system utility routines,
3. 16 KByte video refresh RAM, which stores high resolution graphic display (130,000 dots) or 8 pages of alphanumeric information,
4. 2 RS-232 serial interfaces, with reprogrammable drivers included in the operating system,
5. 2 8-bit parallel general purpose interfaces,
6. 4 programmable timer/counter channels,
7. 2 integrated minidisk drives with 220Kbyte (formatted) each, with double density/double sided versions optional,
8. 9" industry quality standard monitor (monochromatic),
9. Separate operator keyboard (7-bit ASCII code),
10. Separate examinee keyboard, and
11. System expansion bus and slots for 2 S100-bus or 8 europe-single-format cards.

Psychological Section

Whole batteries of tests or inventories are presented automatically, including test instructions and training sessions. There is a selectable mode of presentation for each test. The proctor/psychologist may choose between power, speed, or power-with-time-limit tests. Selection of items to be presented follows a sequential strategy or another type of adaptive strategy. There is a selectable alphanumeric and/or high resolution graphic display of test items on the screen.

Examinee's answers are input by a simple, easy-to-handle keyboard and are automatically checked for correct, incorrect, or illegal response. Response times for each item are recorded and integrated into the set of test data. There is selectable feedback of correct/incorrect responses and elapsed item or test time.

Any latency between the presentation of successive items is eliminated by the "advanced processing concept." All jobs of computation, item selection, and picture processing for the next item are performed during the presentation of the actual item. (There is a delay of 15 seconds maximally for the preprocessing of graphical items, when the button "next item" is pressed continually without reading the item and solving the problem).

Apparatus tests--like reaction time measurement, vigilance tests, or tracking tests--may be integrated into the test battery and controlled by the computer system (real-time execution). The test data are stored on disk after each subtest during the test session, thus avoiding loss of data due to power outage. Test data are integrated into a structured test data bank after each session for further test or item analysis procedures.

Physiological Section

In connection with the Zak-A/D interface, the TEST 2000 system allows on-line recording and analysis of up to six different physiological input channels. Long time off-line recording and analysis of psychophysiological data are accomplished by the pocket-sized, accu-driven microprocessor system BIOPORT, which interfaces directly to the TEST 2000 system. The BIOPORT system is especially suited for concurrent measurement of psychophysiological variables during test or training situations, thus revealing the examinee's reactions and resistance to stress-inducing factors. At present, this section is not implemented in the FAF; however, acquisition is planned for pilot selection and evaluation procedures.

Problems

There are currently problems in two areas. The first problem concerns the data transfer from the microcomputer to a large sized computer for further calculations. At present the data (scores and latencies/time for each subtest) are printed and then manually input to a larger computer. This problem will be solved in the fall of 1983 by using a tape drive and magnetic tape for the data transfer from the microcomputer at Munich to the large computer in Bonn. The second problem is that graphic items require a great deal of effort for programmers, since they are written in Assembler language in connection with a graphic preprocessor. In the future, video disks will be used to store the graphic items; these items will then be monitored by a microcomputer.

Test Material and Software

Aptitude Classification Battery

The first empirical phase started with the Aptitude Classification Battery (EVT), which is the standard test battery for entrance into the Federal Armed Forces and is quite similar to the ASVAB. At present six subtests of the EVT (the subtests without graphic items) are on the computer in two parallel forms (see Table 1). Further, four subtests that are not on the computer (e.g., radio test, test for reaction rate) measure special aspects and are speeded tests.

All six tests on the computer have time limits; within these limits, all items without an answer are presented again. Testing time varies between 45 and 75 minutes, since some examinees need more time for the sample items, whereas others do not use the entire time allocated for a subtest. It is practical to have, on an average, eight persons each day at the test station.

In the first empirical phase, volunteers who have had experience with the EVT battery have been tested. These persons took the EVT as draftees at the recruiting center. When they become volunteer-carriers, they take the EVT once more using a computer-administered parallel form. The next samples of testees will consist of draftees and persons without experience with the EVT and will be grouped according to education level. Furthermore, a small sample will be tested with the two ways of item presentation on the screen (green vs. black).

Software

The software is stored on the system disk. The disk contains, for example,

Table 1
The Six EVT Subtests Used for the FAF Pilot Project

Test	No. of Items/Alternatives	Time
Word Analogy Test (WAT)	20 items/5 alternatives	4 1/2 minutes
Figure Reasoning Test (FDT)	20 items/8 alternatives	10 minutes
Arithmetic Reasoning Test (RT)	20 items/input of the results	14 minutes (paper-and-pencil for notes)
Spelling (Orthographical Test; RST)	50 items/correct-incorrect	3 minutes
Mechanical Ability Test (MT)	20 items/5 alternatives	13 minutes
Electrotechnical Comprehension Test (EKT)	20 items/4 alternatives	10 minutes

the following programs (with content requested by a system command):

ZAKDOS: Operating system of TEST 2000.
EDIT: Editor for input and data management.
BASIC: Interpreter for BASIC (FORTRAN will soon be available).
GRAPHIC: Preprocessor for the graphic items (in Assembler).
SEQUEN: Management of the items with item text and alternatives (e.g., storage on disk and deletion of the blanks).
EVTA1, EVTA2, etc.: Item text and alternatives for the EVT.
EVTB1, EVTB3, etc.: The nongraphic subtests in two parallel forms.
EVT: Management of the EVT test session (creation of the data file, common instructions for EVT, item presentation--conventional/sequential). The item material is stored separately from the presentation procedure, so the system is very flexible.
TESBAS: Users' program in BASIC for handling the data/results, e.g., output on screen or printer, with results for each item (answer, scoring, latency/response time) or for the six subtests (scores, time used).

Results

Empirical Results

Table 2 shows the results calculated from the data of the first 208 examinees. These persons had first taken the EVT by paper and pencil (some weeks or months earlier) and then had taken the EVT by computer at the recruiting center for volunteers in Munich.

The comparison using the *t* test for dependent samples shows significant mean differences except for Arithmetic. The verbal subtests Word Analogy and Spelling had higher scores using paper and pencil and the subtests with figural items (Figure Reasoning, Mechanical, Electrotechnical) had higher scores using

Table 2
Scores from the Six Subtests of the EVT
for the Computerized Adaptive Test
and Paper-and-Pencil Test

Subtest	CAT	Paper and Pencil	t	p	r
Word Analogy	14.56	15.13	-2.86	.005	.606
Figure Reasoning	16.50	15.65	4.81	.000	.703
Arithmetic	12.01	11.70	1.59	.114	.785
Mechanical	13.37	12.58	4.45	.000	.734
Spelling	31.39	34.00	-5.47	.000	.772
Electrotechnical	7.19	6.14	4.11	.000	.682

computerized testing. The correlation coefficients varied between .60 and .79. Further analysis of these data and collection of data are in progress.

Other Results

There have been no problems handling the testing station by the psychologist, subsequent to a one-day training session by Zak. For example, the psychologist, after a one-day training session by Zak, starts the test session prepares the EVT allocating a data file (for data security, code numbers for the examinees are stored on the disk only), and prints the results. An additional aspect of the training was to reduce anxiety/reservation towards the computer and to familiarize the psychologists with the fact that the computer gives some assistance and support.

The examinees have not had any problems handling the test computer using the special keyboard. Before starting the session there is a motivation phase during which the psychologist gives an introduction to the computerized testing session. Then, the examinee is seated in front of the CRT, reads all further instructions on the screen, and uses the keyboard (paper and pencil for the Arithmetic test only for calculations). Questions are rare, except concerning the brightness of the screen.

The volunteers respond positively to this individualized way of testing. The difference between the fastest and slowest examinee is about 30 minutes. There is enough time allocated so that examinee can reread the instructions until they have understood how to solve the items. On the other hand, some examinees do not need the entire time allocated for a subtest and go on to the next subtest when they have answered all items. It is therefore possible to save some testing time with this individualized testing procedure.

There is positive response to feedback from the computer. After each item, the examinee presses the button "next item," and the computer reacts and presents the next item. There is also positive response to the special testing situation. During the paper-and-pencil session, about 50 persons are seated in one room, and the tests are carried out in groups. During the computerized testing session there are only four examinees at the test computer, with the psychologist seated nearby for motivational purposes at the beginning of the testing session and for answering questions during the course of the test.

Future Developments

Hardware

At the end of October 1982 a major change in the hardware will occur. There will be an integrated testing station, which will consist of a microcomputer TEST 2000 for monitoring the testing station, with floppy disks, tape drive, and line printer, and about five testing terminals with a special screen for graphic items and moving pictures (especially developed for tracking items), 64 KByte RAM, and the special keyboard for the examinee. Two testing places will have amber screens, so that it will be possible to compare different forms of item presentation. The system is planned to connect eight testing stations via an interface with TEST 2000. The host computer loads a subtest to the RAM of the testing location, and the examinee then takes this subtest on his/her terminal. Following a subtest the results are stored on the disk, and the next subtest is loaded.

Second Empirical Phase

After the first empirical phase is terminated, the second empirical phase is scheduled for November 1982. Following the conventional testing phase, adaptive testing with fixed branching strategies (pyramidal testing) will commence. Three tests have thus been prepared: Word Analogy, Number Series, and Spelling, each consisting of 66 items and 11 stages. The items were prepared and selected in a pilot study, with the branching procedures developed using classical testing theory and, additionally, Mokken's (1971) probabilistic approach for the check of unidimensionality/homogeneity (Nauels & Wildgrube, 1981). There will thus be an opportunity to compare conventional results (collected either by paper and pencil or by computer) with data collected either by computerized adaptive strategies for the Word Analogy and Spelling tests.

Third Empirical Phase

In the third empirical phase, evaluation of variable branching strategies in adaptive testing are planned. It is still in an uncertain stage, however, because the FAF is searching for testing procedures that can be used in the daily selection process and that, furthermore, can be understood by the test users.

Parallel with starting the second phase of pyramidal testing, the first empirical approach (collection of data) to use the response time of each item (latency) as an additional ability parameter will be carried out (Birke, 1981).

Other Research

In June 1982 research was begun with Lutz Hornke, University of

Duesseldorf. Small projects are being undertaken (1) to investigate the "human factor" aspects of computerized testing (e.g., man-machine interaction, ergonomic aspects such as different screen colors, influence of different instructions for the test procedure) and (2) to compare these results with conventional paper-and-pencil testing. The item pool will be prepared for figural items (similar to the Figure Reasoning test) using logical rules for item construction to vary item difficulty based on Embretson's (1983) work. Collection of data for these items is planned in conventional paper-and-pencil format at the beginning of 1983.

REFERENCES

Birke, W. *Item response time as a basis for ability and difficulty measures*. Paper presented at the 23rd annual conference of the Military Testing Association, Arlington VA, October 1981.

Embretson, S. E. Component latent trait models for test design. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1983.

Mokken, R. J. *A theory and procedure of scale analysis with applications in political research*. New York: deGruijter/Berlin: Mouton, 1971.

Nauels, H.-U., & Wildgrube, W. Probleme des Itempools beim adaptiven Testen--Pilotstudie zum CAT. *Zeitschrift f$r Differentielle und Diagnostische Psychologie*, 1981, *4*, 303-323.

Wildgrube, W. Computerized testing in the Federal Armed Forces. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, Psychometric Methods, 1980.

DESIGN OF A MICROCOMPUTER-BASED ADAPTIVE TESTING SYSTEM

C. DAVID VALE
ASSESSMENT SYSTEMS CORPORATION

Research leading toward the implementation of adaptive testing has been actively underway for over 10 years. Initially, this research focused on developing and evaluating psychometric models and item selection strategies applicable to the adaptive testing process. Now that much of the basic research has been done, work on adaptive testing is moving in two directions. One direction emphasizes the development of increasingly sophisticated latent trait models; the other is directed toward the engineering problems of making the implementation of adaptive testing feasible.

Although the need for a computer system to support adaptive testing has been recognized from the beginning, relatively little emphasis has been placed on the development of a turnkey testing system. The computer programming required to develop, administer, and analyze an adaptive test represents a considerable investment in time and effort. Although a few computerized testing systems have been developed, they have been either small special purpose systems designed to administer and score a single test or large general purpose research systems. A general purpose self-contained adaptive testing system has not yet been developed.

Until recently, a major obstacle to development of such a system was a lack of sufficient demand for such a system to warrant anyone committing the resources necessary for its development and support. A capable computerized adaptive testing (CAT) system is usually somewhat dependent on the computer hardware, and to develop systems for all of the computers in use by adaptive testers would have been very impractical.

The availability of microcomputer systems and components have changed this, however. Capable microcomputers are sufficiently inexpensive that they can be dedicated exclusively to testing. It is now possible to design a testing system around a single computer and still have a system that is cost effective for testing.

The project described in this paper was sponsored by the Office of Naval Research. Its goal was to explore the feasibility of developing a single-user microcomputer-based testing system by (1) considering the needs of the potential consumers, (2) evaluating the available hardware, and (3) designing software that would run on available hardware and meet the needs of the consumers.

Evaluating the Needs of the Consumers

The evaluation of user needs proceeded along two lines. In the first approach, characteristics of current psychological tests were considered. Since the basic element of a test is the test item, a survey of the testing literature was done to discover what types of items might be used in the system. Furthermore, since a testing strategy is important to an adaptive test, the literature was also reviewed to compile a list of all of the strategies that might be used in such a system. In the second approach, potential users were surveyed. Several were interviewed, and a questionnaire was developed and distributed to determine what features they would like to see in the testing system.

Item Types

Five types of items were identified in the literature. The most basic was the knowledge item. A knowledge item is a question that seeks to determine whether the examinee possesses a knowledge of something. The most common of this type of item is the dichotomously scored multiple-choice item. Although some of these items require graphic displays, processing the responses requires only that the response given be matched to a small number of possibilities (e.g., five). Variations of the basic knowledge item include probabilistic response items (deFinneti, 1965), answer-until-correct items (Hanna, 1975), confidence-weighted items (Shuford, Albert, & Massengill, 1966), and free-response items (Vale, 1978). Items of these latter types require somewhat more complex processing of the responses.

The second type, the cognitive process item, assesses whether an examinee possesses a skill without, in theory, requiring the examinee to possess any knowledge. Examples of this type of item were contained in Cory's (1977, 1978) Memory for Objects Test. Items in this test tachistoscopically presented frames containing four to nine objects. The examinee's task was either to recall and to type the names of the objects or to recognize the objects from a word list. Administratively, the cognitive process items add a requirement to precisely time the stimulus; the remaining requirements are similar to those of the knowledge items.

Perceptual motor items comprised the third item type. An example of a perceptual motor item is the computerized version of the Two-Hand Coordination Test described by Hunter (1978). This test requires the examinee to move a cursor to follow a target as it moved around the screen. Items of this type add two additional requirements. First, sufficient processing power is required to manipulate a moving stimulus. Second, an analog input device, such a joystick, may be required.

A fourth type of item was the simulation. Simulations can be as complex as a high fidelity aircraft simulator. Such simulations must be considered beyond the current scope of a microcomputer-based testing machine. Less complex simulations, such as the medical simulation described by Prestwood (1980), are within the capabilities of current microcomputer systems, however. In the simulation described by Prestwood, medical residents were allowed to find symptoms, to administer treatments, and to cure (or kill) patients. This process was implemented as a branched sequence of scenarios with multiple-choice responses. Administratively, items such as this require no more system resources than knowledge items. To be easily implementable, they will require an authoring system that allows the branching among items to be easily specified.

The final item type consisted of noncognitive items. This type includes everything that does not assess a cognitive ability and includes personality items, interest items, and value items. Administratively, these items require no capabilities other than those required by the previous item types.

Testing Models

Most of the tests administered in paper-and-pencil form are conventional linear tests that administer a fixed sequence of items to all examinees. Most strategies of interest for a microcomputer-based testing system are adaptive. Vale (1981) grouped most of the adaptive strategies of interest into three categories: interitem branching strategies, intersubtest branching strategies, and model-based branching strategies.

The interitem branching strategies are implemented by structuring an item pool so that testing begins with one item, and each response to each item leads to the administration of a specific item. In the typical interitem branching strategy, a correct response to one item leads to a more difficult item and an incorrect response leads to a less difficult item. Examples of interitem branching strategies are the pyramidal and Robbins-Monro strategies (Weiss, 1974).

The intersubtest branching strategies are similar in concept to the interitem branching strategies except that the branching is from subtest to subtest rather than from item to item, a subtest being simply a group of items. Vale (1981) further divided this class of strategies into reentrant and nonreentrant forms. A reentrant form is one in which administration can branch to a subtest, out of that subtest to another subtest, and back into the original subtest. A nonreentrant form does not allow a subtest to be exited and then reentered. The two-stage test (Angoff & Huddleston, 1958) is one example of a nonreentrant strategy. Examples of reentrant strategies are Lord's (1971) flexilevel strategy and Weiss's (Vale & Weiss, 1978) stradaptive strategy.

The model-based strategies usually build on item response theory (IRT) and select items and score responses to those items in order to optimize the testing process according to the model. Procedurally, they start with an estimate of ability, select the item that is expected to most refine that estimate of ability, refine the estimate using the response to the item, and repeat the process until the estimate is adequately refined. One popular form of this type of strategy is Owen's (1975) Bayesian strategy. This strategy starts with a Bayesian prior distribution representing an initial estimate of ability. An item is selected that is expected to minimize the variance of the posterior distribution. This item is administered and used to produce the posterior distribution. A popular variant of this strategy selects items based on statistical information. This accomplishes nearly the same result but requires less computation.

For the interitem and intersubtest branching models, the difficult task in system design is to provide a convenient method of specifying the branching algorithms. For the model-based strategies, the greatest challenge is in selecting and designing a set of statistical programs that will allow the testing strategies to be implemented efficiently.

A Survey of Consumers

Twenty-seven individuals from seven different organizations were interviewed to gain further insight into the needs of the tester. In these interviews each individual was given a brief description of what was being developed and was then asked to suggest what features would be particularly useful to his/her application and to suggest what other features not included in the description would be useful. From these interviews, as well as from the literature reviewed, a questionnaire was developed to determine what features potential users would like to see in a microcomputer testing system and what features their tests would require. The questionnaire was divided into four sections. The first asked about characteristics of current tests and tests that might not be implemented on the proposed system. The second section addressed characteristics of the test administration process. The third asked what features would be desirable on a system, considered in light of the cost. The final section asked how much computer experience the respondent had.

Questionnaires were sent to 108 individuals. These individuals were selected for having attended conferences on computerized testing or having written articles in the area. Fifty-five questionnaires were returned, fifty with useful data (the other five individuals declined to respond.)

The test characteristics section of the questionnaire addressed four issues: the size of the item pools, the types of items, the testing strategies, and the types of scoring. Analyses of the questionnaire suggested that to accommodate item pools sufficiently large to satisfy the needs of 75% of the respondents, the system would have to store approximately one million characters just for the items. To accommodate this percentage of respondents, the system would also have to provide a way of timing the item and would have to allow graphics of at least the line-drawing variety. Almost all respondents said that they would administer conventional tests on the system, 68% said they would use model-based adaptive procedures, and 36% said that they would use interitem or intersubtest branching strategies. Seventy percent said that they would use some form of IRT scoring.

The test administration section also addressed four issues: what population would take the tests, whether the administrations would be proctored, what potential problems the system should monitor, and how much abuse the system was likely to get. Although 26% said that the system would be used for a grade-school populations, the most use was anticipated for high school (77%) and college (65%) groups. Eighty-eight percent of the respondents said that a proctor would always be available, 12% said that a proctor would usually be available, and only one respondent said that a proctor would not be available. Most of the respondents felt that the system should be able to detect unreasonably long response delays, random responding, and aberrant responding. On the issue of abuse, most felt that given the oportunity, examinees might push extra buttons and might spill things on the testing terminal; approximately one fourth thought that the examinees might be unnecessarily rough or might inflict deliberate damage on the system.

The system procurement section asked what optional features the respondents would like for the system. More than half of the respondents said that they would like a display capable of black-and-white line drawings, a videodisc or videotape interface, a simplified keybord, and a language (e.g., FORTRAN) compiler.

The final section focused on the computer skills of the potential test developer. Of the 49 individuals who responded to this section, all but two had run computer programs and all but three had written them. Ninety-two percent had written computer programs in a language like BASIC, Pascal, or FORTRAN. When asked whether they would like to develop tests in an author language or using a menu system, one third preferred the author language, one third preferred the menu system, and one third had no preference.

Design of a System to Meet User Needs

The design of a system to meet the needs of the users consisted of (1) selection of the hardware and (2) design of the software. The hardware search has been described by Vale, Albing, Foote-Lennox, and Foote-Lennox (1982) and will not be discussed here, with the exception of two points. First, at the time of the study there were a variety of microcomputers available that were adequately suitable. Second, the typical deficiency that prevented a system from being acceptable was the lack of sufficient disc storage for the item banks.

Existing Software

It is the software that makes a testing system different from any other microcomputer application. Work on the design of this part of the system began with a review of existing testing systems. Unfortunately, very few adaptive testing systems existed, and even fewer of those were documented in the literature. The most relevant system documented in the literature was the TCL system (Vale, 1981). This system was a self-contained test specification and administration system designed for a small single-user microcomputer. The system supported three testing functions: item banking, test development, and test administration. The item banking system was rudimentary; it simply read items from a sequential source file that could be edited by the system text editor and inserted them into a randomly accessible item bank. Tests were specified using an author language: a programming language developed especially to specify adaptive test structures. A test compiler translated the language into an easily executable form, collected the items required by the test, and produced an executable file for administration. The test administrator read the executable test file and administered the test. As much of the time-consuming computation as possible was done during test compilation so that it would execute faster when the test was administered. As an example of this type of computation, items for a maximum information testing strategy were sorted into a table organized by ability level during the compilation. When the test was executed, the time-consuming search was reduced to a fast table-lookup procedure.

The TCL system had some shortcomings in its design, however. Foremost among these was the somewhat clumsy nature of the author language. Although the language provided a means of test specification much simpler than FORTRAN programming, the test specifications were rather difficult to read. Furthermore, the implementation of the language allowed a reference to a specific item to appear only once in the test specification. This made some strategies (e.g., Robbins-Monro) impossible to specify without creating several copies of an item with different reference numbers in the item bank. Finally, the TCL language blurred the distinctions among item reference numbers, variables, functions (i.e., scoring procedures), and statement labels. Although this allowed the skilled user to develop "clever" specifications, it invited trouble when used by a more typical test developer.

Since so few testing systems existed, computerized instructional systems (CAI) were also considered. The most useful design idea came from the Control Data PLATO system. This idea was two-level authoring. For the skilled lesson author, an author language was provided. This language allowed great flexibility in lesson design. For the subject-matter expert who was not greatly skilled in programming, a menu system was available to allow lesson material to be inserted into an existing lesson format. This approach allowed a division of labor between the lesson writers and the programmers.

Design of a Complete Self-Contained System

The most complete testing to date, the TCL system, had three components: item banking, test specification, and test administration. A complete system needs at least one and probably two more components. Since much of adaptive testing is built around the item characteristics, a complete system must include a test analysis system so that, minimally, the important characteristics of the items can be estimated. Additionally, such a system should also include a method for reporting the results to the examinees. The software system designed in this project had all five components.

The complete design is detailed in the report by Vale et al. (1982). From the test developer's point of view, the most interesting part of the design is the test authoring system. Similar to PLATO, the test authoring system was a two-level system. The most flexible and comprehensive level is the author-language level. All tests developed in this system must at some point be specified in the author language. Everything that the system is capable of doing, in terms of test administration, can be controlled through the author language. Because the author language is comprehensive, it can be difficult for an inexperienced test developer to use, especially if she/he has not programmed before. To assist test authors who are not proficient programmers, a menu system forms the second level in the system. This level does not allow much flexibility in test design but it does allow a test developer to construct tests without having to program, even in an author language.

The author language is moderately structured and consists of eight types of statements. A test specification in the language is composed of test modules.

Subtests can be nested within tests (the system makes no formal distinction between a test and a subtest). To allow tests and subtests to be distinguished, two module delimiter statements are provided. The TEST statement denotes the beginning of a test or a subtest module and assigns it a name. The ENDTEST statement denotes the end of a test or a subtest module.

Variables are allowed in the language and are used to hold constants, scores, or other numerical values. In the language, as designed, variables are implemented as words of up to 10 characters. Local variables are defined only within a test module; local variables with the same name in different test modules will be different variables. Global variables are defined throughout all test modules; global variables with the same name in different test modules will be the same variable, and values will thus transfer from one module to another. Global variables are distinguished from local variables by beginning with an "$" sign. The second type of statement, the assignment statement, allows the test developer to assign the results of expressions to variables. The assignment statement in the author language is called SET.

The third type comprises the item statement. This statement, consisting minimally of a "#" sign followed by an item reference number, causes an item to be administered or included in an adaptive structure. It can control, to some extent, how the item is administered and can control what happens after the item is administered.

The declarative statements, SETSCORE and TERMINATE, determine what scores are calculated, what variables are assigned to them, and under what conditions the testing is to terminate. These statements are called declarative because they do not cause anything to happen; they just set up what will happen when the test is administered.

Each administration of a test will create a data file for the examinee. The output control statements, KEEP and AUTOKEEP, determine what is written to this file. KEEP writes specified identifying information and the values of specified variables to the file each time it is executed. AUTOKEEP is a declarative statement that works like KEEP but is set once at the beginning of the test module and thereafter executes automatically each time an item is administered.

The conditional statements, IF, ELSEIF, and ENDIF, allow logical expressions to determine whether a section of a module will be executed. The IF statement begins a logical clause. The ELSEIF statement continues the logical clause until a logical expression is satisfied. The ENDIF statement denotes the end of the logical clause.

The adaptive statements SEARCH, ENDSEARCH, SEQUENCE, and ENDSEQUENCE allow model-based and reentrant adaptive structures to be built. SEARCH executes a maximum information item search based on a specified variable and includes in the search all items listed between the SEARCH and ENDSEARCH statements. The SEQUENCE statement is used for reentrant intersubtest branching strategies. A SEQUENCE statement, similar to a SEARCH statement, uses the items listed between the SEQUENCE and the ENDSEQUENCE statements. The SEQUENCE statement administers one item, starting at the top of the list, every time it is executed. It keeps a pointer so that it does not administer any item more than once. Strata for the stradaptive testing strategy may be set up as sequences.

Finally, the menu statements allow a test template with blanks in it to be written. Menu statements are not really elements of the author language but rather are statements that control a template preprocessor. The INSTRUCT statement causes the preprocessor to print a line of instructions at the test developer's terminal when the preprocessor is run. Blanks (actually underline characters) signify that user input should be read by the preprocessor and inserted into the author language. Qualified blanks limit the kind of information that will be accepted (e.g., item reference numbers) or the number of items that will be accepted. A template containing these statements can be processed by the preprocessor to collect the information needed to complete the test specification from the test developer.

An Example of Test Specification

All test specification is done in the author language. Figure 1 presents an author language specification for a Bayesian adaptive test. This specification consists of a single test module, a test named BAYES. The specification begins by setting up the scores to be kept and the conditions for termination. This test will compute a Bayesian score and will keep the posterior mean and variance in the variables MEAN and VARIANCE. Termination will occur when the posterior variance drops below .01 (or when all of the items have been administered). The prior mean is initialized at 0.0 and the prior variance at 1.0.

Figure 1
Author Language Program for Bayesian Test

```
TEST BAYES ! BAYESIAN ALGEBRA TEST
!
   SETSCORE BAYESIAN (MEAN, VARIANCE) ! SET THINGS UP
   TERMINATE (VARIANCE < 0.01)
   SET MEAN = 0.0
   SET VARIANCE = 1.0
!
   SEARCH MEAN ! FIND ITEM MOST INFORMATIVE AT MEAN
      #MATH001    ! ASSOCIATIVE RULE
      #MATH013    ! COMMUTATIVE RULE
      #MATH144    ! SOLVE FOR X
      *MAPOOL     ! OLD ALGEBRA POOL
      *HARDPL     ! NEW DIFFICULT ALGEBRA POOL
      #MATH552    ! ROOT ITEM
      #MATH603    ! SET ITEM
   ENDSEARCH
!
   KEEP "ALGEBRA ", "MEAN AND", "VARIANCE", MEAN, VARIANCE
!
ENDTEST
```

The items listed between SEARCH AND ENDSEARCH will be considered, as well as the items included in the pools MAPOOL and HARDPL. Each item will be selected to maximize the information for an examinee with an ability level of MEAN, the current ability estimate. When the test ends, the two scores will be written to the data file, along with the label "ALGEBRA MEAN AND VARIANCE."

Although the author language specification for this test is not particularly complex, specification of the same test can be simplified considerably in the menu mode. Figure 2 shows how a test developer would develop the same test using the template preprocessor. The test developer has only to answer the questions as they are asked. The preprocessor produces an author language specification equivalent to the one shown in Figure 1.

Figure 2
Menu Type Test Specification

```
TEST DEVELOPMENT SYSTEM -- TEST COMPLETER PROGRAM      Ver. 1.0

What is the name of the template you would like to use?  BAYES

This template creates an adaptive test based on Owen's Bayesian
algorithm.  It assumes a N(0,1) prior and keeps the final posterior
mean and variance as scores.

Enter a title for the test :  ALGEBRA 101 MIDQUARTER -- 4/82

Now enter the items or pools you would like to include in
the test, one on each line.  Leave a blank line when you are
done.

? #MATH001
? #MATH013
? #MATH144
? *MAPOOL
? *HARDPL
? #MATH552
? #MATH603
? _

Construction of the test is complete.

Under what name would you like to save the new test?  ALG482

The test has been saved as ALG482.  Would you like to create
another?  NO

Test development is then over.  Remember to say BYE to the computer.
```

Figure 3 shows the template that the preprocessor requires to produce the author language equivalent to that in Figure 1 from the interactive session shown in Figure 2. It differs in two respects from the specification in Figure

1. First, it has blanks where item numbers and other constants will go. Second, it has instructions for the test developer that tell what is to be entered when the preprocessor prompts the user for a response. When this template is preprocessed, the blanks are replaced with information supplied by the test developer, and the menu statements are removed.

Figure 3
Template for Bayesian Test

```
! BAYESIAN TEST TEMPLATE
!
INSTRUCT This template creates an adaptive test based on Owen's Bayesian
INSTRUCT algorithm. It assumes a N(0,1) prior and keeps the final posterior
INSTRUCT mean and variance as scores.
INSTRUCT
INSTRUCT Enter a title for the test :
!
TEST BAYES !  __
!
   SETSCORE BAYESIAN (MEAN, VARIANCE) ! SET THINGS UP
   TERMINATE (VARIANCE < 0.01)
   SET MEAN = 0.0
   SET VARIANCE = 1.0
!
INSTRUCT
INSTRUCT Now enter the items or pools you would like to include in
INSTRUCT the test, one on each line. Leave a blank line when you are
INSTRUCT done.
INSTRUCT
!
   SEARCH MEAN ! FIND ITEM MOST INFORMATIVE AT MEAN
   #__#
   ENDSEARCH
!
   KEEP "ALGEBRA ", "MEAN AND", "VARIANCE", MEAN, VARIANCE
!
ENDTEST
!
INSTRUCT
INSTRUCT Construction of the test is complete.
INSTRUCT
```

Current Status of System Development

The design of the CAT system described in this paper is complete. It appears to be a useful design for an operational CAT development and administration. Work on development of the system is currently proceeding in two directions. First, a system very similar in design to the one discussed here is under development for implementation on a Digital Equipment Corporation PDP-11 computer system. This system will be a multi-user testing system incorporating most of the features of the microcomputer design as well as several additional features that were feasible on the larger PDP-11 system. This system is being written in Pascal and will be complete by the end of 1982.

Development of the microcomputer-based system described here has not yet begun. This development is scheduled to begin in mid-1983 as a second phase for the project in which the design was developed. The microcomputer-based system will be developed on a state-of-the-art microcomputer. Prototype systems will be given to test developers to try out, and modifications will be made to the system in order to enhance its utility. A CAT system should thus be available that will make the implementation of an adaptive test as simple as the implementation of a conventional test.

REFERENCES

Angoff, W. H., & Huddleston, E. M. *The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test* (Statistical Report 58-21). Princeton NJ: Educational Testing Service, 1958.

Cory, C. Relative utility of computerized versus paper-and-pencil tests for predicting job performance. *Applied Psychological Measurement*, 1977, *1*, 551-564.

Cory, C. Interactive testing using novel item formats. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

de Finetti, B. Methods for discriminating level of partial knowledge concerning a test item. *The British Journal of Mathematical and Statistical Psychology*, 1965, *18*, 87-123.

Hanna, G. S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, 1975, *12*, 175-178.

Hunter, D. R. Research on computer-based perceptual testing. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, *8*, 147-151.

Owen, R. J. A Bayesian seuential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 1975, *70*, 351-356.

Prestwood, J. S. The development of a computerized, domain-referenced assessment system for family medicine. In R. L. Holloway (Ed.), *Some issues surrounding the implementation of objectives-based education in family medi-*

cine. Minneapolis: University of Minnesota, Department of Family Practice and Community Health, 1980.

Shuford, E. H., Albert, A., & Massengill, H. E. Admissible probability measurement procedures. Psychometrika, 1966, 31, 125-145.

Vale, C. D. Computerized administration of free-response items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Vale, C. D. Design and implementation of a microcomputer-based adaptive testing system. Behavior Research Methods and Instrumentation, 1981, 13, 399-406.

Vale, C. D., Albing, C., Foote-Lennox, L., & Foote-Lennox, T. Development of a microcomputer-based adaptive testing system. St. Paul MN: Assessment Systems Corporation, June 1982.

Vale, C. D., & Weiss, D. J. The stratified adaptive ability test as a tool for personnel selection and placement. TIMS Studies in the Management Sciences, 1978, 8, 135-151.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1974.

ACKNOWLEDGMENT

This research was sponsored by the Office of Naval Research under Contract N00014-82-C-0132 through the Defense Small Business Advanced Technology Program.

DISTRIBUTION LIST

Navy

1 Code N711
Attn: Arthur S. Blaiwes
Naval Training Equipment Center
Orlando, FL 32813

1 Dr. Nick Bond
Office of Naval Research
Liaison Office, Far East
APO San Francisco, CA 96503

1 Dr. Robert Breaux
NAVTRAEQUIPCEN
Code N-095R
Orlando, FL 32813

1 Dr. Robert Carroll
NAVOP 115
Washington , DC 20370

1 Dr. Stanley Collyer
Office of Naval Technology
800 N. Quincy Street
Arlington, VA 22217

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Dr. Charles E. Davis
Personnel and Training Research
Office of Naval Research (Code 442PT)
800 North Quincy Street
Arlington, VA 22217

1 DR. PAT FEDERICO
Code P13
NPRDC
San Diego, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Ms. Rebecca Hetter
Navy Personnel R&D Center (Code 62)
San Diego, CA 92152

Navy

1 Dr. Norman J. Kerr
Chief of Naval Education and Training
Code 00A2
Naval Air Station
Pensacola, FL 32508

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Daryll Lang
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. William L. Maloy (02)
Chief of Naval Education and Training
Naval Air Station
Pensacola, FL 32508

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr William Montague
NPRDC Code 13
San Diego, CA 92152

1 Ms. Kathleen Moreno
Navy Personnel R&D Center (Code 62)
San Diego, CA 92152

1 Commanding Officer
Navy Personnel R&D Center
Code 0C
San Diego, CA 92152

1 Commanding Officer
Navy Personnel R&D Center
Code 06
San Diego, CA 92152

1 Commanding Officer
Navy Personnel R&D Center
Code 41
San Diego, CA 92152

Navy

10 Commanding Officer
Navy Personnel R&D Center
Code 62
San Diego, CA 92152

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

1 Technical Director
Navy Personnel R&D Center
San Diego, CA 92152

6 Personnel & Training Research Group
Code 442PT
Office of Naval Research
Arlington, VA 22217

1 Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

1 Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152

1 Dr. Mary Schratz
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Alfred F. Smode
Senior Scientist
Code 7B
Naval Training Equipment Center
Orlando, FL 32813

1 Dr. Richard Snow
Liaison Scientist
Office of Naval Research
Branch Office, London
Box 39
FPO New York, NY 09510

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

Navy

2 Mr. Brad Sympson
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Frank Vicino
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Ronald Weitzman
Naval Postgraduate School
Department of Administrative Sciences
Monterey, CA 93940

1 Dr. Douglas Wetzel
Code 12
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
San Diego, CA 92152

Marine Corps

1 Col. Ray Leidich
Headquarters, Marine Corps
MPI
Washington, DC 20380

10 Major Frank Yohannan, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

Army

1 Dr. Kent Eaton
Army Research Institute
5001 Eisenhower Blvd.
Alexandria , VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Clessen Martin
Army Research Institute
5001 Eisenhower Blvd.
Alexandria, VA 22333

1 Dr. Karen Mitchell
Army Research Institute
5001 Eisenhower Blvd
Alexandria, VA 22333

1 Dr. William E. Nordbrock
FMC-ADCO Box 25
APO, NY 09710

1 Mr. Robert Ross
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

10 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joyce Shields
Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Hilda Wing
Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Air Force

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Col. Roger Campbell
AF/MPXOA
Pentagon, Room 4E195
Washington, DC 20330

1 Dr. Alfred R. Fregly
AFOSR/NL
Bolling AFB, DC 20332

1 Dr. Patrick Kyllonen
AFHRL/MOE

Brooks AFB, TX 78235

1 Dr. Randolph Park
AFHRL/MOAN
Brooks AFB, TX 78235

1 Dr. Roger Pennell
Air Force Human Resources Laboratory
Lowry AFB, CO 80230

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Maj. Bill Strickland
AF/MPXOA
4E168 Pentagon
Washington, DC 20330

1 Dr. John Tangney
AFOSR/NL
Bolling AFB, DC 20332

1 Major John Welsh
AFHRL/MOAN
Brooks AFB , TX 78223

1 Dr. Joseph Yasatuke
AFHRL/LRT
Lowry AFB, CO 80230

BEST COPY AVAILABLE

Departrent of Defense

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

10 Dr. Anita Lancaster
Accession Policy
OASD/MI&L/MP&FM/AP
Pentagon, Room 2B271
Washington, DC 20301

1 Dr. Jerry Lehnus
OASD (M&RA)
Washington , DC 20301

1 Dr. Clarence McCormick
HQ, MEPCOM
MEPCT-P
2500 Green Bay Road
Noprth Chicago, IL 60064

1 Dr. W. Steve Sellman
Office of the Assistant Secretary
of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

Civilian Agencies

1 Dr. Helèn J. Christup
Office of Personnel R&D
1900 E St., NW
Office of Personnel Management
Washington, DC 20015

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Private Sector

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
Personalstammamt der Bundeswehr
D-5000 Koeln 9C
WEST GERMANY

1 Dr. R. Darrell Bock
Department of Education
University of Chicago
Chicago, IL 60637

1 Mr. Arnold Bohrer
Section of Psychological Research
Caserne Petits Chateau
CRS
1000 Brussels
Belgium

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52243

1 Dr. Ernest R. Cadotte
307 Stokely
University of Tennessee
Knoxville, TN 37916

1 Dr. James Carlson
American College Testing Program
P.O. Bcx 168
Iowa City, IA 52243

1 Dr. John B. Carroll
409 Elliott Rd.
Chapel Hill, NC 27514

Private Sector

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. Hans Crombag
Education Research Center
University of Leyden
Boerhaavelaan 2
2334 EN Leyden
The NETHERLANDS

1 Mr. Timothy Davey
University of Illinhois
Department of Educational Psychology
Urbana, IL 61801

1 Dr. Dattpradad Divgi
Syracuse University
Department of Psychology
Syracuse, NE 33210

1 Dr. Fritz Drasgow
Department of Psychology
University of Illinois
603 E. Daniel St.
Champaign, IL 61820

1 Dr. Stephen Dunbar
Lindquist Center for Measurement
University of Iowa
Iowa City, IA 52242

1 Dr. John M. Eddins
University of Illinois
252 Engineering Research Laboratory
103 South Mathews Street
Urbana, IL 61801

1 Dr. Susan Embertson
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
Lawrence, KS 66045

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Benjamin A. Fairbank, Jr.
Performance Metrics, Inc.
5825 Callaghan
Suite 225
San Antonio, TX 78228

Private Sector

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

1 Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3939 O'Hara Street
PITTSBURGH, PA 15260

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 Dipl. Pad. Michael W. Habon
Universitat Dusseldorf
Erziehungswissenshaftliches Inst. II
Universitatsstr. 1
D-4000 Dusseldorf 1
WEST GERMANY

1 Dr. Ron Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

1 Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

1 Prof. Lutz F. Hornke
Universitat Dusseldorf
Erziehungswissenschaftliches Inst. II
Universitatsstr. 1
Dusseldorf 1
WEST GERMANY

1 Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 90010

Private Sector

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
603 East Daniel Street
Champaign, IL 61820

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Douglas H. Jones
Advanced Statistical Technologies
Corporation
10 Trafalgar Court
Lawrenceville, NJ 08148

1 Dr. William Koch
University of Texas-Austin
Measurement and Evaluation Center
Austin, TX 78703

1 Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Robert Lockman
Center for Naval Analysis
200 North Beauregard St.
Alexandria, VA 22311

1 Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Private Sector

1 Dr. James Lumsden
Departrent of Psychology
University of Western Australia
Nedlands W.A. 6009
AUSTRALIA

1 Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

1 Mr. Robert McKinley
University of Toledo
Dept of Educational Psychology
Toledo, OH 43606

1 Dr. Barbara Means
Human Resources Research Organization
300 North Washington
Alexandria, VA 22314

1 Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

1 Dr. W. Alan Nicewander
University of Oklahoma
Departrent of Psychology
Oklahoma City, OK 73069

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. James Olson
WICAT, Inc.
1875 South State Street
Orem, UT 84057

1 Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Cirle, NW
Washington, DC 20036

1 Dr. James Paulson
Dept. of Psychology
Portland State Unive3rsity
P.O. Box 751
Portland, OR 97207

Private Sector

1 Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 Lowell Schoer
Psychological & Quantitative
Foundations
College of Education
University of Iowa
Iowa City, IA 52242

1 Dr. William Sims
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 Martha Stocking
Educational Testing Service
Princeton, NJ 08541

1 Dr. Peter Stoloff
Center for Naval Analysis
200 North Beauregard Street
Alexandria, VA 22311

1 Dr. William Stout
University of Illinois
Department of Mathematics
Urbana, IL 61801

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Kikumi Tatsuoka
Computer Based Education Research Lab
252 Engineering Research Laboratory
Urbana, IL 61801

Private Sector

1 Dr. Maurice Tatsuoka
220 Education Bldg
1310 S. Sixth St.
Champaign, IL 61820

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Mr. Gary Thomasson
University of Illinois
Department of Educational Psychology
Champaign, IL 61820

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. Ledyard Tucker
University of Illinois
Department of Psychology
603 E. Daniel Street
Champaign, IL 61820

1 Dr. David Vale
Assessment Systems Corporation
2233 University Avenue
Suite 310
St. Paul, MN 55114

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. Ming-Mei Wang
Lindquist Center for Measurement
University of Iowa
Iowa City , IA 52242

1 Dr. Brian Waters
HumRRO
300 North Washington
Alexandria, VA 22314

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

Private Sector

1 Dr. Rand R. Wilcox
University of Southern California
Department of Psychology
Los Angeles, CA 90007

1 German Military Representative
ATTN: Wolfgang Wildegrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington , DC 20016

1 Dr. Bruce Williams
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

1 Ms. Marilyn Wingersky
Educational Testing Service
Princeton, NJ 08541

1 Dr. George Wong
Biostatistics Laboratory
Memorial Sloan-Kettering Cancer Center
1275 York Avenue
New York, NY 10021

1 Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

18. SUBJECT TERMS (continued)

response-contingent testing, item characteristic curve theory, latent trait test theory, item response theory

19. ABSTRACT (continued)

3. Multidimensional Item Response Theory. Paper by Roderick P. McDonald; discussion by Fumiko Samejima. Paper by Mark D. Reckase and Robert L. McKinley; discussion by Fritz Drasgow.

4. Estimating Parameters with the E-M Algorithm. Papers by Robert K. Tsutakawa, Robert J. Mislevy and R. Darrell Bock; discussion by Charles Lewis.

5. Unidimensionality and Robustness. Papers by William Stout, Fritz Drasgow and Charles K. Parsons, Douglas H. Jones; discussion by David Thissen.

6. Adaptive and Sequential Testing. Papers by David J. Weiss and Debra Suhadolnik, R. A. Weitzman; discussion by Mark D. Reckase.

7. Latent Trait Models for Special Applications. Paper by Susan Embretson (Whitely); discussion by Douglas H. Jones. Paper by Kikumi K. Tatsuoka; discussion by Susan Embretson (Whitely).

8. Applications of Computerized Adaptive Testing. Papers by James R. McBride and J. B. Sympson, and Loralee Hartmann, Wolfgang Wildgrube, C. David Vale.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Governmen.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ONR 85-1	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION University of Minnesota 6b. OFFICE SYMBOL *(If applicable)*	7a. NAME OF MONITORING ORGANIZATION Personnel and Training Research Programs Office of Naval Research
6c. ADDRESS *(City, State, and ZIP Code)* Department of Psychology Minneapolis, MN 55455	7b. ADDRESS *(City, State, and ZIP Code)* Arlington, VA 22217-5000
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL *(If applicable)*	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-82-G-0061
8c. ADDRESS *(City, State, and ZIP Code)*	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
61153N	RR042-04	042-04-01	150-487

11. TITLE *(Include Security Classification)*
Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference

12. PERSONAL AUTHOR(S)
David J. Weiss

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Year, Month, Day)*	15. PAGE COUNT
FINAL REPORT	FROM 82JUL27 TO 82JUL30	1985, April	371

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Ability testing, achievement testing, sequential testing, automated testing, branched testing, individualized testing, tailored testing, programmed testing, test theory, (cont.)

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report is the Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference held July 27-30, 1982 at the Spring Hill Conference in Wayzata, Minnesota. These Proceedings include all papers presented at the conference and discussions of these papers by the scheduled discussants.

The papers are organized into the following sessions:

1. Developments in Latent Trait Theory. Paper by Fumiko Samejima; discussion by Roderick P. McDonald. Paper by Michael V. Levine; discussion by Robert J. Mislevy.

2. Parameter Estimation. Papers by Frederic M. Lord and Marilyn S. Wingersky, David Thissen and Howard Wainer; discussion by Michael V. Levine. Paper by Charles Lewis; discussion by Robert K. Tsutakawa.

(continued on the other side)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE *(Include Area Code)*	22c. OFFICE SYMBOL
Dr. Charles E. Davis	(202) 696-4046	Code 442PT

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED